

Université de Marne-la-Vallée,

THÈSE
pour obtenir le grade de
Docteur de l'Université de Marne-la-Vallée

présentée et soutenue publiquement par M. Mario Monteleone

8 décembre 2003

LEXICOGRAPHIE ET DICTIONNAIRES
ELECTRONIQUES
DES USAGES LINGUISTIQUES AUX BASES DE DONNÉES
LEXICALES

LEXICOGRAPHY AND ELECTRONIC
DICTIONARIES
FROM LINGUSITIC USES TO LEXICAL DATA BASES

Directeurs de thèse: M. Elia Annibale, M. Laporte Éric

M.me Mirella Conenna, professeur à l'Université de Bari (Italie), rapporteur

M.me Mireille Piot, professeur à l'Université Grenoble 3 (France), rapporteur

TABLE DES MATIÈRES

TABLE DES MATIÈRES	2
AVANT-PROPOS	3
0. INTRODUCTION	4
0.1 AFFRONTER ET VAINCRE LE MINOTAURE	4
0.2 LES ESPACES INFINIS ET MAGNETIQUES DE L'ELECTRONIQUE	6
0.3 C'EST LE LEXIQUE OU LA MORT!.....	10
CHAPITRE I. DELIMITATION ET PROBLEMATIQUE DE L'ETUDE: JANUS A DOUBLE VISAGE	15
1.1 LE BEAU MODÈLE ANTIQUE	17
1.2 LA BASE DE DONNEES LEXICALE	22
1.3 LE MOTEUR LINGUISTIQUE	25
1.3.1 <i>La crise inattendue de Monsieur George Boole</i>	26
1.3.2 <i>La correction privée des spelling checkers</i>	43
1.4 LA "SEPARATION DOLOUREUSE"	47
CHAPITRE II. ANATOMIE D'UN DICTIONNAIRE ELECTRONIQUE	53
2.1 LE LEXIQUE ETIQUETE.....	55
2.2 LE DICTIONNAIRE FLECHI.....	71
2.3 LE DICTIONNAIRE DE MOTS COMPOSES	78
2.4 MOTS, SIGNIFIES ET USAGES DANS LES DICTIONNAIRES ELECTRONIQUES.....	81
2.4.1 <i>Comment choisir les lemmes</i>	82
2.4.2 <i>La sémantique dans le dictionnaire électronique</i>	87
CHAPITRE III. DICTIONNAIRE ELECTRONIQUE VERSUS DICTIONNAIRE PAPIER: LE ROI EST NU!	93
3.1 ABSENCES INJUSTIFIEES	94
3.1.1 <i>L'affaire Perniola</i>	110
3.1.2 <i>Freud versus Popper, i.e. la Psychanalyse versus le Faillibilisme</i>	113
3.1.2 <i>Quelle fin ont eu les verbes pronominaux?</i>	120
3.1.1.1 Le système Intex® en italien	122
3.1.1.2 Les agglutinations de l'italien	135
3.1.1.3 Automates à états finis et reconnaissance automatique des textes	142
3.1.1.4 Quelques considérations sur les verbes pronominaux	160
3.2. LA GUERRE ENTRE LES LEXIQUES	165
3.2.1 <i>Pronoms ou numéraux?</i>	166
3.2.2 <i>Countables et Uncountables</i>	171
3.2.3 <i>Le lexique en fonction du lexique</i>	176
POUR NE PAS CONCLURE	183
BIBLIOGRAPHIE	188

AVANT-PROPOS

Ce travail résume ce qui a été un long processus de croissance personnelle, culturelle et scientifique, qui a duré plus de dix ans et pendant lequel j'ai eu la chance de rencontrer les plus importants experts, pas seulement italiens, du sujet que je m'apprête à traiter. En les mentionnant comme les acteurs d'un film, en ordre d'apparition, je veux donc saisir l'occasion favorable pour les remercier tous, et en particulier Mirella Conenna, professeur à l'Université de Bari, maître et amie, première à dévoiler les mécanismes du langage naturel; Jacques Labelle, professeur à l'Université du Québec à Montréal; Annibale Elia, professeur au Dipartimento di Scienze della Comunicazione de l'Université de Salerne, personne qui représente aujourd'hui pour moi un véritable point de repère humain et scientifique; Simona Vietri, et Emilio D'Agostino, eux aussi professeurs au Dipartimento di Scienze della Comunicazione de l'Université de Salerne; Mireille Piot, professeur à l'Université Grenoble 3, Christian Leclère, chercheur du L.A.D.L de l'Université de Marne-la-Vallée, Eric Laporte, professeur à l'Université de Marne-la-Vallée, Max Silberztein, professeur à l'Université de Franche-Comté, créateur du logiciel Intex®, et Giustino De Bueriis, davantage qu'un ami, chercheur au Dipartimento di Scienze della Comunicazione de l'Université de Salerne. Les mots, les discours et les réflexions échangés avec toutes les personnes susmentionnées ont été pour moi illuminantes, et m'ont aidé à me former dans celle qui est devenue mon activité principale.

Mais parmi toutes les personnes qui ont indirectement participé à la rédaction de cet écrit, il y en a une à laquelle vont mes plus profonds remerciements, malheureusement posthumes. Cette œuvre est ainsi dédiée à Maurice Gross, l'un des plus importants linguistes du XXème siècle, disparu le 8 décembre 2001, et envers lequel la linguistique formelle et informatique reste en dette éternelle pour tout ce que qu'il a su comprendre par intuition et créer. Si le souvenir de l'ami et le grand héritage méthodologique et scientifique du Maître ne disparaîtront jamais, il me reste néanmoins le grand regret que Maurice Gross ne puisse aujourd'hui lire les pages qui suivent. Elles n'auraient rien pu lui dire qu'il ne sût déjà, mais pour moi cela aurait été un véritable honneur de recevoir ses conseils et ses remarques.

0. INTRODUCTION

0.1 AFFRONTER ET VAINCRE LE MINOTAURE

“Le langage est un labyrinthe de routes. On vient d’un côté, et on arrive à s’orienter; on parvient au même endroit d’un autre côté, et on ne s’y retrouve plus.”¹

Le mythe nous dit que le Labyrinthe de Cnossos, créé par Dédale et habité par le Minotaure, avait une conformation très complexe, raison pour laquelle tous ceux qu’y entraient ne pouvaient en sortir sans aide, et surtout sans tomber sous les griffes du Monstre. Afin de soustraire Ariane au sacrifice auquel elle avait été destinée, Thésée décidait de s’introduire dans le Labyrinthe et de tuer le Minotaure, en déroulant une longue pelote de laine dont Ariane tenait à l’extérieur l’extrémité. Après avoir affronté et vaincu le fils de Pasithae, le héros grec utilisa le fil déroulé pour retrouver la porte de sortie.

Lévi-Strauss nous a expliqué que tous les mythes sont dépourvus de valeur pratique et ont comme objectif celui de résoudre logiquement une contradiction, de façon à consentir à l’intelligence humaine de définir et d’ordonner hiérarchiquement les phénomènes qu’elle rencontre. Le sens logique du mythe d’Ariane et de Thésée est déjà manifeste dans le sens du terme Labyrinthe, d’origine préhellénique, et qui implique le concept du chemin dur, accidenté, impossible à parcourir sans un fil conducteur ou un guide.

Comme tous les mythes, celui-ci a donc aussi une valeur universelle, qu’on pourrait résumer ainsi: tous ceux qui s’apprêtent à accomplir un projet, comme nous dans les pages qui suivent, vont rencontrer inmanquablement quelque type de labyrinthe ou de Minotaure, c’est-à-dire des difficultés, des problèmes imprévisibles et des fautes possibles. Parcourir indemnes le chemin et vaincre le Monstre signifie dépasser tous les pièges et rejoindre leur but sans commettre d’erreurs. Pour faire cela, il

¹ Wittgenstein, L., 1995, page 109 (c’est nous qui traduisons).

est nécessaire d'avoir une méthode de recherche établie à laquelle se référer, c'est-à-dire qu'il faut suivre un fil conducteur, justement comme Ariane et Thésée.

En partant de ces prémisses, à l'aide des pages qui suivent, nous allons donc entrer dans le *labyrinthe de routes* préconisé par Wittgenstein, et nous confronterons deux disciplines très différentes mais qui s'intéressent toutes deux au langage naturel: la lexicographie, qui a une longue tradition, et la linguistique-informatique, qui est beaucoup plus récente. En particulier, nous mettrons en relation deux *produits* spécifiques de ces disciplines, i.e. les dictionnaires papier pour la lexicographie et les dictionnaires électroniques pour la linguistique-informatique, dans le but de détecter différences et similitudes et de vérifier quel est en réalité entre les deux le serviteur le plus fidèle et le meilleur miroir de l'usage de la langue italienne. Pour nos argumentations, nous utiliserons des outils spécifiquement linguistiques, mais aussi logiques et liés au sens commun, et à la fin, nous chercherons à organiser les résultats de notre recherche en une sorte de *grammaire*, dont les règles puissent servir soit à la réalisation des dictionnaires papier, soit à la création des dictionnaires électroniques.

Nous avons défini notre *labyrinthe* et aussi le fil conducteur que nous allons suivre, mais il est difficile de dire si, à travers notre analyse, nous arriverons à vaincre le Minotaure. A l'état actuel, nous ne relevons pas de fautes de cohérence sérieuse dans ce que nous allons exposer, et nous laissons à nos *adversaires* la tâche de chercher des erreurs et des contradictions. Nous considérerons les critiques les plus féroces comme des témoignages d'estime et d'attention, et si notre exposition trouve tout le monde d'accord unanimement, nous saurons avoir reformulé de façon plus explicite des solutions à quelques questions. Et alors, nous serons prêts pour partir à la recherche de nouveaux labyrinthes à parcourir et de nouveaux minotaures à affronter.

0.2 LES ESPACES INFINIS ET MAGNETIQUES DE L'ELECTRONIQUE

Pendant la dernière décennie, l'informatique et les instruments qu'elle utilise se sont emparés rapidement et d'une façon pénétrante de la vie quotidienne de beaucoup de personnes, et aujourd'hui il semble impossible de rencontrer quelqu'un qui n'ait encore connu les nombreux avantages de l'informatisation, phénomène qui a déterminé des changements radicaux et irréversibles pour presque toutes les activités humaines. L'importance de ce que nous pouvons définir comme "l'offensive informatique" nous apparaît dans sa totalité si nous réfléchissons à l'un des plus profonds changements qu'elle a fait naître, i.e. la révolution ou la virtualisation des concepts d'espace, temps et transportabilité, surtout en ce qui concerne la communication et la diffusion du savoir. Internet, "le réseau des réseaux", et avec celui-ci le courrier électronique, les *chat-lines*, les vidéoconférences, ont réduit la distance entre les locuteurs, et ont accéléré tous ces phénomènes qui sont consécutifs à l'échange d'informations. L'espace, du point de vue de la distance physique, a donc été presque annulé par la technologie. Aussi, les supports laser tels que les CD-Roms et les DVD-Roms, légers, facilement transportables et avec de grandes capacités de stockage, permettent d'organiser et d'archiver le savoir sur la base de modalités révolutionnaires, aussi bien du point de vue qualitatif que quantitatif.

En même temps, la rencontre entre l'informatique et la linguistique a donné naissance à la linguistique-informatique, discipline qui se situe à cheval sur les sciences humaines et les sciences humaines exactes et qui s'occupe de *"l'étude des systèmes d'élaboration consacrés à la compréhension et à la génération du langage"*². D'ailleurs, la rencontre entre ces deux disciplines n'a pas été accidentelle, étant donné que historiquement l'informatique s'est toujours occupée du langage naturel, en principe pour fonder et puis perfectionner les mécanismes et les interfaces qui permettent aux "humains" d'utiliser le langage naturel afin de dialoguer de façon interactive avec les ordinateurs. De son côté, la linguistique a trouvé dans l'informatique beaucoup de supports, puissants et hétérogènes, qui se sont révélés nécessaires pour développer les

² Grishman, R., 1988 (c'est nous qui traduisons).

simulations d'opérations cognitives complexes, pour vérifier pratiquement les présuppositions théoriques descriptives que la linguistique elle-même élabore, ou aussi, plus simplement, pour analyser automatiquement et en une seule fois des corpus de grande envergure, une activité qui dans le passé aurait été matériellement impossible. Dans ce dernier cas, et surtout récemment, les capacités de stockage des supports magnétiques et optiques comme les disques durs de nouvelle génération, les CD-Roms gravables et regravables, les DVD-Roms se sont démontrées de grande importance. Elles permettent de sauvegarder et de réutiliser aisément de grandes quantités de données, en facilitant leur classification, l'analyse et la gestion automatique.

Dans son domaine, pendant des années, la linguistique-informatique a conduit à différentes expérimentations, surtout celles qui, grâce aux ordinateurs, simulent en langage naturel les plus diverses et complexes activités, comme la traduction automatique et assistée par l'ordinateur, l'analyse textuelle automatique, le parsing³, la reconnaissance et la génération automatique des textes.

³ En ce qui concerne le domaine du traitement automatique des langues, le parsing peut être défini comme le processus d'assignation de descriptions structurelles aux séquences de mots produites par le langage naturel (ou aux séquences de symboles dérivées par des séquences de mots). Le type de description structurelle à assigner et les modalités d'assignation dépendent de la grammaire – c'est-à-dire d'un langage descriptif et d'un ensemble de restrictions structurelles – sur la base de laquelle le parser essaye d'analyser les séquences de symboles qu'on lui soumet. Autrement dit, un parser prend comme input une séquence de mots (ou leur subrogé) en une langue donnée et une description abstraite des relations structurelles possibles qui peuvent exister entre les mots ou les séquences de mots de cette langue, et donne comme output zéro ou plusieurs descriptions structurelles de l'input, sur la base de ce qui est prévu dans l'ensemble des règles structurelles. On aura zéro description si la séquence d'input ne peut pas être analysée par la grammaire (i.e. si elle n'est pas grammaticale ou si le parser est incomplet, c'est-à-dire s'il n'arrive pas à trouver toutes les structures possibles pour la grammaire). Par contre, on aura plus d'une seule description si l'input est ambigu par rapport à la grammaire, c'est-à-dire si la grammaire prévoit plus d'une analyse correcte de l'input.

Pour un parser, la séquence de symboles de l'input peut aussi n'être pas formée seulement par des mots en langage naturel. En ne considérant pas le parsing de langages artificiels (comme ceux de la programmation ou de la logique), de documents étiquetés (par exemple en SGML) ou de séquences non linguistiques telles que celles des codes génétiques, le parsing dans le traitement du langage naturel peut être effectué sur des séquences de mots, d'étiquettes de parties du discours ou sur des séquences de symboles complexes telles que les additions de caractéristiques (i.e. là où un mot peut être substitué par un ensemble de caractéristiques, y compris sa forme orthographique, la partie du discours, la classe de flexion et ainsi de suite).

En général, on effectue un parsing parce qu'on estime que les structures grammaticales d'une séquence participent au signifié et que localiser la structure grammaticale d'une séquence de mots en langage naturel est une démarche nécessaire pour l'individuation du sens de cette séquence. Avec un certain type de parser, la construction d'une représentation du sens est effectuée simultanément à la dérivation d'une analyse sur la base de la grammaire.

Les activités et les applications de cette discipline se sont développées et multipliées ultérieurement, et ont contribué à redéfinir ses objectifs et ses fonctions:

“linguistique-informatique

L'étude du 'langage naturel à l'aide de l'ordinateur. Même si dans les faits les recherches de linguistique-informatique sont souvent liées à celles de l'intelligence artificielle, il est d'usage de distinguer entre linguistique-informatique et élaboration (automatique) du langage naturel (ELN) parce que la première ne poursuit pas la réalisation de systèmes artificiels capables de prestations intelligentes par rapport au langage, mais la connaissance du langage même; souvent, elle utilise l'ordinateur comme instrument de vérification de théories linguistiques indépendantes. En outre, font partie de la linguistique-informatique (mais pas de l'ELN) les recherches qui utilisent des techniques informatiques "non intelligentes", telles que celles de la stylistique informatique et celles basées sur l'élaboration (également à l'aide d'instruments statistiques) de corpus lexicaux, en vue de la réalisation de vocabulaires, concordances, et ainsi de suite. La linguistique-informatique est d'autre part active dans tous les secteurs de la recherche linguistique théorique, de la syntaxe à la pragmatique et à l'analyse du discours, avec la construction de systèmes qui réalisent des théories ou bien des fragments de théories linguistiques".⁴

Cette distinction entre la linguistique-informatique et l'ELN représentera l'un des points basiques pour ce que nous exposerons dans la suite. C'est dans la visée de la *connaissance du langage même* que nous voulons classer les arguments dont nous parlerons, en démontrant que *l'élaboration de corpus lexicaux* est en réalité une

Les opérations de parsing vont du simple isolement syntagmatique, par exemple avec le but de reconnaître les noms propres, à l'analyse sémantique d'un texte, pour la récupération d'informations ou la traduction assistée par l'ordinateur.

⁴ Beccaria G. L. (éd.), 1994 (c'est nous qui traduisons).

opération fondamentale dans les recherches de linguistique-informatique. Nous observerons en fait que sans la réalisation et l'analyse de ces corpus il ne serait pas possible d'accéder à une étude exhaustive de la syntaxe – mais aussi de la morphologie – ni évidemment à l'*analyse du discours*, pour ne reprendre que deux des points mentionnés.

En partant de ces bases, nous verrons qu'une étude taxinomique du lexique et de ses propriétés grammaticales représente la démarche essentielle pour approcher des concepts et des activités tels que l'analyse textuelle automatique, le parsing et les applications d'automates à états finis.

0.3 C'EST LE LEXIQUE OU LA MORT!

Dans un lumineux chapitre de l'un de ses livres les plus récents⁵, Steven Pinker réfute de façon irréversible la désormais centenaire hypothèse du relativisme et déterminisme linguistique, mieux connue comme Hypothèse Sapir-Whorf, en soutenant avec des démonstrations presque incontestables qu'il existe une nette séparation entre pensée et langage, et que le premier étant du point de vue cognitif antécédent au deuxième, le langage ne peut en aucune façon modeler la pensée des êtres vivants, quelle que soit la langue qu'ils parlent. La pensée, identifiée comme modalité de représentation subjective de la réalité ou comme élément expressif de l'imagination, peut donc exister même en l'absence, ou carence, d'activités linguistiques et communicatives, tandis qu'il est assez difficile de démontrer l'opposé.









La thèse de Pinker est extrêmement fascinante et réellement sensée. De notre côté, nous voulons toutefois ici réaffirmer en même temps que le langage naturel ne pourrait exister sans la capacité, typiquement humaine, d'assigner des noms aux choses, et d'utiliser ces noms à l'intérieur d'activités communicatives articulées – même avec la prétention de ne pas être mal compris. Pour cela, nous utiliserons en premier un cliché bien connu: on dit habituellement que le jeu est une activité très sérieuse, étant donné qu'en jouant, on peut comprendre des faits et événements également très profonds et complexes. Nous allons donc vérifier encore une fois cette affirmation en proposant de lire le texte:

(1)

⁵ Pinker, S., 1994.

On se demandera si ce qui précède est réellement un texte, et la question paraît légitime. En réalité, il s'agit d'une phrase brève dont les caractères alphabétiques ont été automatiquement convertis par un logiciel de vidéo-écriture dans un set de caractères non alphabétiques. Chacun des symboles correspond donc à une lettre de l'alphabet italien, et chaque séquence graphique correspond à un mot. En termes plus techniques, un texte semblable est dit crypté et possède une clé spécifique, à appliquer pour son décryptage. Dans notre cas, la clé sera établie par les correspondances (arbitraires) entre symboles et lettres de l'alphabet. Donc, en décodant le texte (1), nous obtiendrons ce qui suit:

(2) *Nessuna lingua al mondo può sussistere senza un proprio vocabolario, un proprio sistema semiotico arbitrario e strutturato nonché un numero di parlanti che la condivide come codice linguistico comunicativo.*⁶

Ecrite de cette façon, notre phrase apparaît clairement, mais les différences entre (1) et (2) ne se réduisent au simple fait qu'en (1) la phrase est cryptée et en (2) elle ne l'est pas. Quant au second texte, par exemple, il nous paraît transparent parce qu'il renvoie à un système linguistique bien défini – la langue italienne – en se conformant à ses canons morphosyntaxiques et sémiologiques. Par contre, le premier "texte" n'est écrit en aucune langue du monde et peut vivre seulement à partir d'un signifié "indirect" en relation avec (2). Les séquences de symboles graphiques en (1), comme par exemple   ? ?   ✓, ne font donc partie d'aucun lexique, et les éléments qui les composent n'entrent dans aucun alphabet graphique et/ou phonétique. Il s'ensuit que pour des séquences comme   ? ?   ✓ il n'est pas possible de repérer un dictionnaire qui

⁶ "Aucune langue au monde ne peut exister sans son propre vocabulaire, son propre système sémiotique arbitraire et structuré ainsi qu'un nombre de locuteurs qui la partage comme code linguistique communicatif" (c'est nous qui traduisons).

nous en explique les signifiés et les usages, et il est donc impossible de leur associer un sens, une prononciation ou une fonction grammaticale.

Ferdinand de Saussure⁷ nous a expliqué que les mots sont des moyens arbitraires créés par les hommes pour définir et désigner les objets, concepts, événements. Luis Godart⁸ a ajouté que l'exigence de désigner et de mémoriser par des formes d'écriture était déjà ressentie plus de trois mille ans avant la naissance de Jésus-Christ. Donc, la création du lexique, conçue comme nécessité d'enregistrer et de garder les rapports entre les signes linguistiques et la réalité connaissable et intelligible, ne peut se vérifier qu'au moment même ou tout de suite après le besoin de dénoter et de connoter le monde environnant. Par conséquent, le lexique doit être ancien comme les langues, même si nous savons qu'en Italie, par exemple, les premiers dictionnaires paraissent seulement en même temps que les Langues Vulgaires. Avant cette époque, en l'absence de dictionnaires, il est possible de supposer que l'usage des mots, c'est-à-dire la préservation des rapports entre signifiants et signifiés, se soit transmis surtout oralement, et seulement de façon limitée par des textes écrits, avec les grandes civilisations comme celles du Moyen Orient, de l'Asie Mineure et de la Méditerranée, notamment les civilisations Grecque et Latine.

Aujourd'hui, ces conditions ont beaucoup changé. L'oralité s'est fortement enrichie grâce aux mass media comme la radio et la télévision, mais elle a partiellement perdu la fonction de transfert narratif du savoir. Aujourd'hui le lexique est transmis et expliqué dans les oeuvres sur papier, parce que l'écriture aussi, y compris l'écriture électronique, a atteint des niveaux très élevés de diffusion.

Une langue sans lexique, pourrait-elle donc exister aujourd'hui? La question nous semble franchement absurde. Si nous mettions à zéro le lexique de l'Italien, serions-nous encore à même de parler et d'exprimer verbalement des concepts? Et si oui, pour combien de temps réussirions-nous à nous souvenir du rapport qui existe entre le mot *herméneutique* et le concept qu'il veut exprimer? En mettant à zéro un lexique, ne sentirions-nous pas immédiatement l'exigence d'en créer un autre?

⁷ Voir Saussure, F. de, 1916.

⁸ Voir Godart, L., 1992.

La réponse à ces questions est probablement unique et même autévidente, i.e. il n'est pas possible d'avoir une langue sans un lexique. Avec les fonctions spécifiques d'ensemble de mots d'une même langue, un lexique structuré a en effet la tâche de fournir aux locuteurs tous les éléments utiles pour réaliser et utiliser les liens entre *denotata* et dénotants. Et si, par absurde, les mots n'avaient pas des référents spécifiques, nous ne pourrions exprimer des concepts ni indiquer des objets et des événements. Si le mot *Nessuna* n'était attesté dans aucun lexique ou dans aucune langue, nous ne pourrions la comprendre et l'utiliser correctement, ou mieux la distinguer de sa séquence cryptée ☹️📦??🤖🔴✓. Et nous ne pourrions même assimiler le sens global de la phrase (2) si tous les autres mots qu'elle contient n'étaient pas attestés. Donc, il n'y aurait aucune différence entre (1) et (2), si les mots du deuxième texte n'étaient signifiants d'aucun signifié: autant ☹️📦??🤖🔴✓ que *Nessuna* seront pour nous de simples séquences graphiques, dépourvues de tout signifié, et cela indépendamment du *mentalese*, le langage mental humain dont Steven Pinker parle dans son livre. Si la pensée existe antérieurement et séparément du langage, ce dernier essaye au moins d'exprimer la pensée, et pour faire cela il utilise des mots. Sans mot, donc, il n'est pas possible d'essayer de manifester la pensée; et renoncer ou négliger l'étude du lexique signifierait aussi mettre au second plan l'instinct du langage de l'homme ou la faculté de langage, ainsi que Ferdinand de Saussure l'a baptisée à l'aube du siècle dernier.

Le jeu que nous avons entrepris au début de ce chapitre nous a amené à définir un aspect essentiel des études sur la communication verbale: vu que le langage ne peut exister sans lexique, l'analyse descriptive de ce dernier, discerné comme matière structurée ou structurable, devient un facteur prioritaire et basique pour toutes les spéculations qui ont pour objet le langage naturel. Parce que le lexique, d'ailleurs, n'est pas une matière inerte mais dynamique, qui change sans cesse, et surtout parce qu'il est le premier "récepteur" de la grammaire d'une langue, vu que les mots, au moins en italien et dans beaucoup de langues romanes, peuvent être subdivisés et fléchis en genre et en nombre, et doivent donc être utilisés en respectant les règles d'accord grammatical. Le signifié des mots, en outre, impose dans l'acte verbal l'application de règles de co-

occurrence et restriction de sélection, qui nous permettent de raccorder un prédicat spécifique à ces compléments, qui sont utiles pour créer des phrases à sens accompli⁹.

Si nous voulons simuler les activités linguistiques humaines à l'aide des ordinateurs, il faut tout d'abord effectuer une étude du lexique approfondie, empirique et taxinomique, sur les résultats de laquelle on pourra ensuite structurer tout autre type d'expérimentation ou d'approche, qu'il s'agisse d'activités de pertinence: soit de l'ELN, soit de la linguistique-informatique. En fait, comme il a déjà été affirmé et démontré par Maurice Gross et le lexique-grammaire, c'est seulement en analysant à fond le lexique qu'il sera possible de comprendre les règles qui gouvernent la combinaison des mots et d'avoir une vision plus claire des aspects morphosyntaxiques, sémantiques et pragmatiques d'une langue donnée. Parce que l'accès conscient à l'emploi du langage naturel ne pourrait se vérifier correctement sans un usage aussi conscient du système référentiel des signes linguistiques, qui est institué par le lexique et qui dans le lexique trouve sa première et fondamentale explicitation.

Cela dit, nous pouvons maintenant tirer quelques conclusions profitables aux questions que nous voulons traiter. D'un point de vue lexicographique le dictionnaire papier, le lieu où la matière lexicale est répertoriée et expliquée, prend pour nous une importance fondamentale, parce qu'il devient un constant point d'arrivée pour l'étude des usages d'une langue, et de plus il est le "réceptacle idéal" de tous les changements et innovations. Par contre, d'un point de vue linguistique, la formalisation du lexique est la démarche la plus essentielle pour la réalisation d'applications plus complexes, comme celles dont nous avons déjà parlé. L'élément crucial et critique pour la linguistique-informatique est donc l'élaboration par l'ordinateur de bases de données lexicales, que nous appellerons dorénavant *dictionnaires électroniques* et qui peuvent avoir, comme nous le verrons, une typologie multiforme¹⁰.

⁹ En effet, ce sont justement ces règles qui nous évitent de produire des phrases comme **Max a bu une pierre*, qui sont amusantes mais qui en fait ne sont pas grammaticales.

¹⁰ Voir Courtois, B. et Silberstein, M. (éds.), 1990.

CHAPITRE I. DELIMITATION ET PROBLEMATIQUE DE L'ETUDE: JANUS A DOUBLE VISAGE

Qu'est-ce exactement un dictionnaire électronique? Quel usage peut-on en faire ? Quels usages pourrait-on en faire ? Et surtout, combien de types de dictionnaires électroniques existe-il? Dans ce chapitre, nous chercherons à donner des réponses plausibles à ces questions, en utilisant des exemples pratiques, sur la base du cadre théorique établi par le lexique-grammaire, et aussi en nous référant aux activités que de nombreuses personnes font quotidiennement avec un ordinateur.

D'un point de vue théorique, un dictionnaire électronique peut être comparé à Janus au double visage, avec un visage tourné vers la linguistique-informatique, qui reste son principal domaine d'application, et l'autre tourné vers la lexicographie, de laquelle il naît comme relecture particulière des buts du dictionnaire papier. Une définition exhaustive de ce type d'ouvrage doit donc tenir en compte cette ambivalence. Pour définir un dictionnaire électronique nous dirons que:

- comme type particulier de dictionnaire, il partage, avec celui sur papier, les finalités de catégorisation et description morpho-grammaticale du lexique. Par conséquent, pendant la phase initiale de sa structuration, le dictionnaire électronique a comme modèle le dictionnaire papier, surtout en ce qui concerne le listage des lemmes utilisés dans une langue donnée;
- d'un point de vue informatique et du point de vu de la linguistique-informatique, c'est une structure informative, comme nous le verrons, réalisée de façon aussi homogène que les bases de données, dont il fait partie;
- en naissant comme la réponse à des exigences spécifiques de la linguistique-informatique, pendant une phase consécutive de structuration, le dictionnaire

électronique s'écarte nécessairement de celui sur papier, pour réorganiser les finalités communes en perspective taxinomique, surtout en ce qui concerne la description morpho-grammaticale du lexique d'une langue donnée.

Nous développerons ces trois points dans les paragraphes qui suivent. Nous traiterons aussi d'autres questions, strictement liées à l'usage des dictionnaires électroniques même si parfois, en apparence, extrêmement différentes de celles d'une base de données lexicales, comme l'est un dictionnaire électronique. Nous verrons par exemple la façon dont des types particuliers de dictionnaires électroniques, créés avec des modalités formelles et informatiques particulières, sont responsables de la faillite de beaucoup de navigations qu'il est possible d'effectuer sur Internet à l'aide des moteurs de recherche Web.

1.1 LE BEAU MODÈLE ANTIQUE

En Italie, la naissance de la lexicographie est inscrite à l'intérieur d'événements vastes et complexes tels que le furent l'origine et le développement des parlers vulgaires, qui pendant les XIII^e et XIV^e siècles, deviennent le moyen de communication de la classe commerciale et assument lentement une importance croissante du point de vue politique aussi bien que littéraire : ce que démontrent les oeuvres de Dante Alighieri. Le *Cantico delle creature* (aux environs de 1225) de Saint François d'Assise, légitime par contre les potentialités évangélisatrices de la langue vulgaire, en créant les premiers éléments d'opposition avec le latin ecclésiastique.

Lentement, l'usage des parlers vulgaires introduit dans la péninsule italienne, les premiers éléments de diglossie¹¹, qui, par conséquent, véhiculent vers l'écriture les premiers vocabulaires et glossaires bilingues, comme par exemple le vocabulaire latin-sicilien de 1348 intitulé *Liber Delcari*, et un glossaire latin-bergamasque autour de 1420.

Pendant l'Humanisme et la Renaissance, le parler vulgaire de la Toscane acquière la condition définitive de nouvelle langue d'usage, au détriment du latin et malgré la forte opposition de l'Eglise qui, en 1577, interdit la lecture de la Bible en langue vulgaire. C'est en 1612 qu'est imprimé le premier dictionnaire moderne monolingue de l'italien, le *Vocabolario della Crusca*. Aujourd'hui, celui-ci représente encore une étape fondamentale de la lexicographie italienne étant donné que (comme le souligne Giovanni Nencioni dans la présentation à la réédition de 1987) "*en ce qui concerne la méthode, le Vocabolario della Crusca marque en 1612, pour le sens historique et le critère systématique, un progrès remarquable par rapport aux dictionnaires compilés au XVI^e siècle: la technique lexicographique fut définie avec beaucoup de soin, comme le démontre l'introduction rigoureuse, et on fit précéder la technique d'une théorie de la langue longuement débattue et expérimentée, qui conféra*

¹¹ Comme le signale Beccaria 1994, on appelle *diglossie* un phénomène de caractère exclusivement sociolinguistique, concernant l'emploi fonctionnellement diversifié des différents codes linguistiques ou des différentes variantes d'un code linguistique à l'intérieur de la même communauté.

à l'œuvre cohésion et caractère"¹². A ces observations, fait écho Beccaria, en soulignant que "*Le Vocabolario degli Accademici della Crusca constitua ensuite le modèle pour toute la lexicographie monolingue européenne*"¹³. Aussi bien avec cette première édition de 1612, qu'avec celles qui suivirent, le *Vocabolario della Crusca* a en fait représenté, pendant des siècles, un point de référence pour les spécialistes et les écrivains de beaucoup de nationalités, et parmi eux les italiens Foscolo, Leopardi et Manzoni.

Pendant les cinq derniers siècles, le dictionnaire monolingue sur papier s'est rarement éloigné du modèle défini en 1612, même en revêtant des formes et des contenus différents. En général, aujourd'hui un dictionnaire papier peut être défini comme une *«œuvre de consultation dans laquelle on décrit le lexique d'une langue parmi une série d'articles composés d'un lemme et d'une glose qui contient une série d'informations sur le lemme (...) L'ensemble des lemmes d'un dictionnaire forme sa liste d'articles ou sa macrostructure, l'article lexicographique en soi constitue sa microstructure (...) Une typologie des dictionnaires se base soit sur les objectifs pour lesquels l'œuvre est compilée et consultée, soit sur la base de différences dans l'organisation des articles ou de la microstructure (...) En général, les lemmes sont ordonnés alphabétiquement (...) Un dictionnaire monolingue présente une seule liste d'articles où les lemmes sont associés à des gloses écrites dans la même langue que le lemme (...) un dictionnaire monolingue générique de 120000 lemmes contiendra un nombre remarquable de termes de spécialité extraits des domaines majeurs du savoir, et aussi des archaïsmes présents chez les auteurs les plus étudiés (...)*¹⁴.

La définition que nous venons de lire est très importante pour la suite de notre exposé, parce qu'elle contient les deux principaux points de contact, pour la forme et le contenu, entre le dictionnaire papier et le dictionnaire électronique, dont nous allons traiter, c'est-à-dire:

¹² de' Rossi, B. (a cura di) 1987 (c'est nous qui traduisons).

¹³ Beccaria G. L., ouvrage cité, page 427 (c'est nous qui traduisons).

¹⁴ Beccaria G. L., ouvrage cité.

- la description du lexique d'une langue donnée;
- l'ordre alphabétique des mots.

L'objectif décrit au premier point, certainement le plus important des deux, est atteint par les deux types de dictionnaire avec des modalités extrêmement différentes, comme nous le verrons par la suite. Nous avançons, pour le moment, que dans notre dictionnaire électronique il n'y aura pas de gloses descriptives, qui seront remplacées par des codes alphanumériques pour la catégorisation grammaticale et flexionnelle. Mais par la suite, nous verrons que, de ce point de vue, la taille des différences existantes entre les deux types de dictionnaires est remarquable: au sujet des modalités descriptives d'un lexique, les méthodes peuvent être plutôt disparates et souvent strictement connexes aux finalités – de l'édition, culturelles, pratiques ou encore commerciales – que se type d'ouvrage poursuit.

Le deuxième point représente presque une exigence: pour le dictionnaire papier, elle est imposée par la nécessité de donner aux lecteurs une méthode de consultation rationnelle, rapide et universellement connue, et pour le dictionnaire électronique par la logique des ordinateurs qui ne permet de traiter automatiquement que les bases de données qui sont ordonnées alphabétiquement.

A partir de la citation précédente, il est nécessaire de souligner un autre aspect important relatif à la *typologie des dictionnaires*, qui est définissable sur la base des *buts pour lesquels l'œuvre est compilée et consultée*. On peut distinguer entre les principaux types suivants de dictionnaires papier:

- monolingues, i.e. les dictionnaires qui décrivent le lexique d'une seule langue;

- monolingues encyclopédiques, dans lesquels les gloses de chaque entrée, à part les informations morphologiques et/ou grammaticales, donnent aussi des descriptions encyclopédiques plus ou moins détaillées, selon la nature de la taille de l'ouvrage;
- bilingues ou plurilingues directionnels, i.e. les dictionnaires qui décrivent le lexique d'une langue source en fournissant pour les lemmes de celle-ci les gloses et les traductions en une ou plusieurs langues cible. Parmi ces dictionnaires, on trouve les ouvrages de spécialités technico-scientifiques, dans lesquels on présente le lexique d'un domaine de la connaissance humaine particulier – par exemple, la physique nucléaire ou la géodésie – avec les traductions correspondantes en une ou plusieurs langues. Souvent, les dictionnaires de spécialités technico-scientifiques ont l'anglais comme langue source;
- bilingues bidirectionnels, i.e. des dictionnaires dans lesquels du point de vue de la traduction, deux langues données peuvent servir alternativement de source dans une section et de cible dans l'autre, et alors on ne donne l'explication de leurs lexiques que quand elles ont une fonction source;

Des versions sur CD-Rom des dictionnaires cités dans la typologie précédente, ont commencé à être commercialisés plutôt récemment par plusieurs sociétés éditrices italiennes et étrangères. Il s'agit de versions électroniques des ouvrages sur papier, enrichies par des outils multimédias de consultation, telles que par exemple la prononciation des lemmes ou quelques structures hyper-textuelles de navigation. De cette façon, il a été créé un nouveau type de dictionnaire, également dit électronique parce que consultable seulement sur ordinateur, mais que nous préférons appeler informatisé, pour le distinguer de la base de données lexicales que nous analyserons mieux par la suite. La distinction est, à notre avis, presque obligatoire, parce qu'un dictionnaire informatisé, à cause de sa structure, ne peut pas, par exemple, être utilisé comme base de données lexicales dans des routines d'analyse textuelle automatique; et

aussi, comme nous le verrons, pour la création de notre dictionnaire électronique, la transposition sur support magnétique ou optique d'un ouvrage sur papier est à la fois superflue et réduite. Pour ces raisons, il sera donc préférable de garder bien séparés les dictionnaires informatisés des dictionnaires électroniques, et nous verrons que cette mise au point terminologique essentielle sera justifiée par de nombreux aspects, de caractère bien théorique qu'applicatif.

1.2 LA BASE DE DONNEES LEXICALE

Qu'est ce qu'une base de données lexicales et quand donc un fichier que nous élaborons peut-il le devenir?

L'informatique établit qu'une base de données est un "ensemble d'informations exhaustives et non redondantes, nécessaires à une série d'applications automatisées et connues par un ensemble logique qui en garantit la gestion. (...) La formulation d'une base de données (...) est un processus qui, en partant de l'observation d'une situation réelle, atteint à une définition de la base de données correspondante"¹⁵. Une base de données est en outre un "ensemble, également très étendu, d'informations de différents types qui font référence à un secteur spécifique du savoir ou à une organisation déterminée. Ces données sont organisées selon des critères précis et par des structures informatives spécifiques, de façon à être consultées, mises à jour et éventuellement restructurées à l'aide de procédures organisées de façon unitaire"¹⁶.

En termes algébriques, les bases de données sont des ensembles finis, puisqu'elles incluent des éléments avec de telles caractéristiques communes qu'on peut les décrire en utilisant une seule méthode. La réalisation d'une base de données est gouvernée par de rigoureux critères d'organisation formelle, qui imposent le catalogage des contenus sur la base de champs et d'étiquettes univoques et non-ambiguës, à appliquer à tous les éléments d'ensemble que l'on veut décrire.

C'est à ce type de formulation qu'aboutit le traitement des bases de données par l'ordinateur, avec la création et l'application de *Data Base Management Systems (DBMS)* spécifiques qui permettent d'accomplir rapidement des opérations de consultation, *l'information retrieval* (récupération d'information), l'importation et l'exportation vers d'autres bases de données, la mise à jour et la copie. En partant des données sur papier, un exemple de base de données pourrait être la transposition sur support optique ou magnétique du contenu des vieilles archives d'état civil d'une mairie, ordonné alphabétiquement et subdivisé sur la base de champs différents tels que les

¹⁵ Morvan, P. 1989, page 36 (c'est nous qui traduisons).

¹⁶ Barcellona, N., Marini, A., Monti, P., Vercesi, M. 1988, page 106 (c'est nous qui traduisons).

noms, les prénoms, les dates de naissance, les adresses et ainsi de suite. En ce cas, nous aurons créé une base de données anagraphique.

Une même rigueur formelle de structuration est appliquée pour la création des dictionnaires électroniques, et sur la base des définitions précédentes, il est possible d'affirmer qu'un dictionnaire électronique est une base de données, vu que:

- il fait référence à un secteur spécifique du savoir, en ce qui concerne le langage naturel et plus précisément le lexique d'une langue;
- sa structuration est faite à partir d'observations spécifiques d'aspects et de phénomènes réels, i.e. ceux relatifs à l'usage d'un lexique;
- pour ce type d'ouvrage, il est possible d'établir des critères homogènes de structuration, vu que les éléments qu'il inclut sont les unités lexicales d'une même langue, qui ont par défaut des caractéristiques en commun;
- il est possible de le repérer exclusivement sur des supports magnétiques et il est utilisé par des systèmes logiques de gestion, c'est-à-dire par des logiciels d'analyse linguistique.

Un dictionnaire électronique est, de fait, une base de données lexicale, comme nous le verrons, avec des fonctions de moteur à l'intérieur des logiciels linguistiques. Néanmoins, l'emploi d'un moteur linguistique est également présent dans d'autres outils informatiques, tels que, par exemple, les moteurs de recherche sur le Web, et nous verrons par la suite jusqu'à quel point la création d'une base de données incomplète ou

son usage approximatif, peuvent annuler ou réduire les possibilités de recherche et de récupération d'information sur Internet.

1.3 LE MOTEUR LINGUISTIQUE

En reprenant les définitions données à propos de la linguistique-informatique, on peut affirmer qu'un dictionnaire électronique ou, mieux, une base de données lexicale fait partie de ces outils informatiques non intelligents, utilisés pour les études sur le langage naturel et la vérification de théories linguistiques indépendantes. Il est toutefois possible d'étendre le champ d'application d'un tel dictionnaire à d'autres activités aussi, pas strictement connexes à des études et des vérifications linguistiques, surtout parce que, comme on l'a déjà dit, un dictionnaire électronique est principalement un outil informatique. Prenons par exemple en considération l'informatisation de l'écriture, c'est-à-dire l'habitude d'aujourd'hui d'écrire ou transcrire sur ordinateur désormais tous les textes dans leur typologie complexe. L'informatisation de l'écriture présente de nombreux avantages et aspects positifs, dont ne sont pas les moindres la rapidité de catalogage, de repérage et de reproductibilité des textes, et aussi leur conservation durable¹⁷, mais elle a aussi fait naître de nouveaux problèmes de gestion et de réception des contenus textuels, tels que le contrôle et la correction automatiques des coquilles, les index de lisibilité, l'analyse statistique et la récupération des informations. L'importance de ces problèmes est énorme, surtout si l'on pense aux documents officiels et de grande divulgation, tels que les textes législatifs, fiscaux ou politiques ou aux ouvrages d'édition plus classiques, tels que les romans, les encyclopédies ou les mêmes dictionnaires papier, et aussi à la myriade de pages Web publiées chaque jour sur le réseau Internet. Ces dernières, en effet sont incontestablement les documents les plus lus et les plus analysés des temps modernes, mais beaucoup d'entre nous ont déjà pu noter qu'au moyen de ces pages, et d'Internet plus généralement, il est souvent très difficile de récupérer rapidement et avec précision les informations nécessaires, même à l'aide des moteurs de recherche les plus sophistiqués possibles et quoique ces informations soient sans doute quelque part, là, dans le réseau Internet.

¹⁷ En ce sens, il est suffisant de penser aux documents gravés sur CD-Rom, des supports qui, comme on le sait, ont un seuil minimal de conservation d'au moins cent ans.

1.3.1 La crise inattendue de Monsieur George Boole

A propos d'Internet, il est possible de donner beaucoup de définitions, mais comme contenant des documents, le réseau des réseaux représente essentiellement des archives très utiles et riches, dans lesquels il est en effet possible de repérer tout type d'information; et même si cette spécificité est strictement dépendante de facteurs difficilement prédictibles, comme par exemple la possibilité d'atteindre un site, l'efficacité ou encore la surcharge des connections téléphoniques et/ou de réseau, la quantité et la qualité des informations normalement contenues dans le Web sont tellement élevées que celui-ci est maintenant devenu un instrument presque irremplaçable.

Avoir des nouvelles de haute qualité est sans doute un avantage, non seulement pour les navigateurs du Web. Mais cette condition peut changer drastiquement si la quantité de nouvelles devient tellement démesurée qu'elle empêche la consultation, au lieu de la favoriser, et en ce cas l'avantage devient rapidement un désavantage. En réalité, c'est le risque qu'on court presque toujours avec les actuelles modalités de recherche disponibles sur le réseau Internet.

Nous savons qu'aujourd'hui, pour localiser ou récupérer du Web une information spécifique, il faut s'en remettre aux moteurs de recherche, c'est-à-dire à des logiciels en ligne qui, à partir de notre recherche, effectuée en utilisant un ou plusieurs mots clefs, lisent les pages présentes sur les différents sites du réseau Internet et nous transmettent les URL des pages qui contiennent le mot ou les mots que nous avons cherchés. Donc, si nous voulions demander toutes les informations contenues dans le Web en relation avec le mot, et donc avec le concept de *fondamentalisme*, en théorie il faudrait insérer ce mot dans la boîte de texte spéciale d'un moteur de recherche, attendre l'élaboration et finalement consulter les résultats qui nous seraient montrés sur l'écran de notre ordinateur.

Ainsi décrit, ce mécanisme de recherche pourrait sembler infallible, mais la réalité est bien différente. Souvent, et surtout avec les moteurs de recherche, la création d'un logiciel doit être rapportée à des questions ou des données qui ne rentrent pas dans

le domaine spécifique de l'informatique ou de la programmation ; et, pour en faire un usage correct, il est nécessaire d'avoir une profonde connaissance des disciplines auxquelles ces questions ou données affèrent, du moins si l'on veut éviter des erreurs, des dysfonctionnements et de fausses attentes. Dans le cas des moteurs du Web, ceux-ci étant des logiciels qui analysent automatiquement des données textuelles, la discipline non-informatique de référence est sûrement la linguistique-informatique, et il faudrait tenir compte des aspects théoriques, pratiques, et aussi des thèmes et activités de celle-ci, lorsqu'on réalise un moteur de recherche. Néanmoins, à cause de ce qu'on pourrait appeler une insuffisante sensibilité linguistique, en général les moteurs de recherche actuels se révèlent peu fiables précisément par rapport à l'activité qu'ils affirment vouloir accomplir, c'est-à-dire : récupérer l'information du réseau Internet.

Mais très exactement, comment fonctionne un moteur de recherche lorsque nous lui soumettons l'une de nos interrogations textuelles? Certes, il ne lit pas complètement et spécifiquement toutes les pages Web du réseau Internet, parce que pour faire cela un moteur emploierait une quantité de temps plutôt élevée et il ne serait pas capable de nous fournir rapidement des résultats. En réalité, les recherches de tous les moteurs sont effectuées à partir d'index des pages Web, donc à partir de fichiers dans lesquels on liste, sans répétitions et en ordre alphabétique, les mots contenus par les différentes pages. La réalisation d'index similaires, qui est appelée *indexation textuelle automatique* et qui est une application ultérieure non-intelligente de la linguistique-informatique, abrège énormément le temps de lecture d'un texte, et permet aux moteurs de recherche de fournir rapidement les résultats.

Une recherche effectuée avec un mot simple, comme par exemple celle indiquée précédemment pour le terme *fondamentalisme*, sera donc en mesure de nous dire avec précision dans quels documents Web le mot est utilisé. Au contraire, la situation change si on fait une recherche avec plusieurs mots car, en ce cas, les moteurs de recherche montrent leurs limites. On sait que le mot *fondamentalisme* exprime des concepts différents que la séquence de mots *fondamentalisme islamique*, et nous pouvons en dire autant pour *carte* et *carte de crédit*. Ces différences, qui sont non seulement formelles mais aussi et surtout sémantiques, sont cependant mises à zéro dans les index réalisés à partir des pages du Web, parce qu'à chaque fois qu'ils apparaissent l'un à côté de l'autre,

les trois mots *carte de crédit*, qui écrits en une séquence ont un signifié unique, sont indexés séparément, c'est-à-dire sont placés sur trois lignes différentes de l'index final, et non sur une même ligne, comme l'exigerait la logique du signifié. La même chose se vérifie pour *fondamentalisme islamique*, *cheval de Troie*, *zone d'ombre*, i.e. pour toutes ces séquences qui sont dites *mots composés* dans le Lexique-grammaire parce que formées par deux ou plusieurs mots mais avec des fonctions grammaticales et sémantiques uniques et, comme nous le verrons, un rôle fondamental pour l'*information retrieval*.

La perte d'informations qui se vérifie à cause de cette typologie d'indexation particulière a des répercussions considérables sur les résultats des recherches effectuées avec les moteurs du Web. En fait, en utilisant l'un quelconque de ceux-ci pour une interrogation simple relativement à *carte de crédit*, la recherche sera effectuée non seulement pour repérer les occurrences contiguës et continues des trois mots, i.e. pour les trois mots écrits l'un après l'autre, mais aussi pour leurs occurrences non-contiguës et discontinuës, i.e. pour les trois mots non écrits l'un après l'autre. Une recherche d'échantillon sur le mot composé italien *carta di credito* (carte de crédit) effectuée au hasard avec un moteur de recherche, a en fait repéré presque 97.700 URL à consulter, et parmi celles que nous avons pu examiner, pas toutes contenaient la séquence exacte *carta di credito*, mais seulement les trois mots éparpillés dans le texte. Ce nombre élevé de résultats représente un véritable obstacle pour la récupération d'informations désirées, parce que la consultation de presque 97.700 URL serait très longue et surtout elle serait brusquement arrêtée par tout navigateur qui se retrouverait lisant même une seule page qui, avec l'information requise, n'aurait que peu à voir¹⁸.

¹⁸ À ce propos, pour un site la recherche de visibilité sur le Web est indirectement l'une des raisons pour lesquelles les résultats des *recherches* sur le réseau Internet sont non réellement pertinentes quant aux informations que l'on désire récupérer. Ce type de visibilité se mesure sur la base des visiteurs qu'un site arrive à accueillir pendant une période de temps spécifique. Normalement, plus élevé est le nombre de visiteurs et plus haute est la visibilité d'un site donné. Avoir une grande visibilité permet aux Webmasters de vendre à prix favorables les espaces publicitaires du site qu'ils ont réalisé, et c'est pour cette raison que la visibilité devient de fait une valeur monétisable.

En ce sens, avoir trop d'informations à consulter peut signifier n'en avoir aucune. Comme on l'a déjà indiqué, et démontré, l'excessive quantité d'informations repérées sur le réseau Internet avec les moteurs de recherche peut parfois ne pas être un avantage.

D'autres problèmes peuvent s'ajouter à ceux provoqués par une indexation incorrecte. En premier lieu, les moteurs de recherche n'indexent pas tout le texte des pages Web, mais seulement une partie, par exemple 30%, et cette partie à indexer peut encore ne pas être séquentielle. En outre, il semble que, pour les textes du Web, il n'y ait pas de procédures standardisées d'indexation, mais que chaque moteur de recherche applique des modalités et des pourcentages différents, et cela fait qu'en présence d'une même recherche, deux moteurs donnent deux résultats différents. En revenant à l'exemple de *carta di credito*, les presque 97.700 URL trouvées par l'interrogation d'un moteur sont devenues 16.101.295 avec un autre, un nombre encore plus inconfortable.

Il faut toutefois dire que les mêmes créateurs et administrateurs de moteurs de recherche doivent s'être aperçu de cet *imbroglio*. En fait, on se souviendra sûrement des premières versions de ces moteurs Web qui avaient des syntaxes de recherche complexes, basées sur l'emploi des opérateurs booléens, ainsi appelés du nom de

Généralement, pour arriver à sortir de l'anonymat, plusieurs Webmasters adoptent des stratégies plutôt particulières, qui créent des désavantages aux navigateurs et rendent, si possible, encore moins crédible la recherche à l'aide des moteurs du Web. Il faut en fait dire qu'après l'avoir créé, les Webmasters doivent signaler l'existence de leur site à un ou plusieurs moteurs de recherche afin que ceux-ci, en présence d'une interrogation quelconque, puissent le consulter pour récupérer les informations requises. A partir de ce moment, les moteurs de recherche auxquels le site est signalé peuvent effectuer l'indexation des pages qu'il contient, mais pour faciliter la récupération des informations est demandé aux Webmasters d'indiquer un ou plusieurs mots-clefs avec lesquels synthétiser le contenu du site même. Pour accroître les possibilités d'être insérés dans les résultats de plusieurs recherches, beaucoup de Webmasters décident d'indiquer des mots-clefs qui n'ont pas de relations pertinentes avec le contenu du site, mais qui sont largement utilisés pendant cette période historique spécifique sur le Web, pour des raisons qui peuvent être de caractère social, culturel, politique et ainsi de suite. Donc, il pourra arriver qu'un nouveau site dédié à la gastronomie signale son existence aux moteurs de recherche en utilisant comme mot-clé par exemple *Usama Bin Laden*, et pour cette raison les navigateurs, en recherchant les dernières nouvelles de politique internationale, se retrouvent à lire des recettes de nouvelle cuisine.

George Boole¹⁹, mathématicien et logicien anglais du XIXe siècle qui postula les lois de leur fonctionnement.

Boole publia presque cinquante écrits, dans lesquels il étudia, entre autres, les propriétés de base des chiffres, de l'arithmétique et de l'algèbre, et aussi la propriété distributive. En ce qui concerne l'étude de la logique, Boole chercha à utiliser dans ce domaine la notation et les principales opérations algébriques. Leibniz avait déjà noté la façon dont il était possible d'associer la disjonction et la conjonction entre autres concepts aux opérations d'addition et multiplication mathématiques, Boole développa ultérieurement et de façon rigoureuse cette intuition, en créant une représentation algébrique non seulement pour la conjonction et la disjonction, mais aussi pour les quantificateurs *tous* et *quelque*, et en utilisant les normes variables algébriques x , y et ainsi de suite comme des variables sur les deux valeurs de vérité, i.e. *vrai* ou *faux*, représentées respectivement par le numéro 1 et le numéro 0. Ce domaine d'étude, dit algèbre de la logique ou algèbre booléenne, porta Boole à la représentation de diverses opérations logiques d'une façon pas très éloignée de celle utilisée par les circuits logiques des ordinateurs d'aujourd'hui.

¹⁹ Né à Lincoln en 1815 et mort à Ballintemple en 1864, George Boole s'occupe de calcul différentiel et élabore une méthode générale de calcul des probabilités; les travaux pour lesquels il est le plus connu, toutefois, sont ceux élaborés dans le domaine de la dite algèbre de la logique. Boole reçoit de son père sa première formation mathématique, et la passion pour la construction des instruments optiques. Ensuite, il étend ses études en se dédiant à la réflexion sur les langues classiques. Il commence à travailler très jeune comme instituteur assistant et à partir de 1835 il ouvre sa propre école, en recommençant à étudier la mathématique en autodidacte. Pendant cette période, Boole étudie les ouvrages de Laplace et Lagrange et est encouragé dans ses études par Duncan Gregory qui est à Cambridge éditeur du "Cambridge Mathematical Journal" qui venait d'être fondé. Même en ne pouvant pas suivre pour des raisons économiques le conseil de Gregory de continuer les cours à Cambridge, Boole commence à publier ses articles sur le "Cambridge Mathematical Journal" et à étudier l'algèbre. Il publie une application de la méthode algébrique pour la résolution des équations différentielles sur les "Transactions of the Royal Society" et pour cet ouvrage il reçoit la médaille de la Royal Society. En 1844 est publié le bref mais fondamental volume de la "Mathematical Analysis of Logic", qui contient une première exposition des concepts qui auraient été recueillis dans une forme plus ample et articulée dans le traité de 1854 *An Investigation into the Laws of Thought, on Which are Founded the Mathematical Theories of Logic and Probabilities* (Une recherche sur les lois de la pensée sur lesquels sont fondées les théories mathématiques de la logique et de la probabilité). Pendant ce temps, en 1849, à Boole est confiée la chaire de mathématique au Queen's College de Cork où il enseigne pour le restant de sa vie. L'étude des équations différentielles est le sujet du *Treatise on Differential Equations* (Traité sur les équations différentielles) de 1859, tandis que le calcul des différences finies est au centre du *Treatise on the Calculus of Finite Differences* (Traité sur le calcul des différences finies) de 1860. Pour ses recherches, Boole reçoit beaucoup d'honneurs, et en 1857 il est élu membre de la Royal Society. Sa carrière s'interrompt tragiquement avec sa mort à seulement 49 ans.

Jusqu'à quelques années auparavant, l'étude des *opérateurs booléens* entraînait dans le domaine exclusif de la théorie mathématique, mais aujourd'hui une connaissance sommaire de leurs définitions et de leurs fonctions est fondamentale pour rendre moins dispersée la consultation du réseau Internet. Quand nous fournissons à un moteur de recherche un terme à localiser, le logiciel construit une *proposition logique* du type "le mot X est présent", la compare avec tous les documents enregistrés dans sa base de données et donne à l'utilisateur seulement les adresses des fichiers pour lesquels elle est vérifiée. Ce processus est dit *test de vérité* et a des résultats univoques s'il est effectué sur un seul mot-clé. Dans le cas où l'utilisateur en insérerait deux, par exemple *géographie physique*, le *test* peut prendre des formes différentes, et un usage correct des opérateurs booléens sert justement à adresser le moteur de recherche vers l'une d'elles en particulier.

Les trois opérateurs booléens fondamentaux sont **AND**, **OR** et **NOT**. L'opérateur **AND** spécifie que **toutes** les propositions logiques construites à partir de la requête de l'utilisateur doivent être vraies pour un document donné, pour que celui-ci soit inséré dans la liste des résultats: par rapport à la séquence précédente, *géographie physique*, le moteur devra donner une réponse affirmative soit à la question "Est-il vrai qu'en ce document le mot *géographie* est présent?", soit à la question "Est-il vrai qu'en ce document le mot *physique* est présent?".

L'opérateur **OR** spécifie qu'au moins l'une des propositions logiques construites à partir de la requête de l'utilisateur doit être vraie pour un document afin qu'il soit inséré dans la liste des résultats: celle-ci indiquera donc tous les sites qui contiennent *géographie* et ceux qui contiennent *physique*, y compris ceux où les deux termes sont présents.

L'opérateur **NOT** spécifie que la proposition logique construite à partir de la requête de l'utilisateur doit ne pas être vraie pour un document afin qu'il soit inséré dans la liste des résultats. Il peut être utilisé pour des recherches à mot-clé simple, mais il n'est pas particulièrement utile de savoir dans quels lieux le terme *géographie* n'est pas présent; s'il est combiné avec d'autres opérateurs, il peut donner des résultats intéressants. Par exemple, la syntaxe logique *géographie AND NOT physique* induira le

logiciel à sélectionner les documents qui contiennent le terme *géographie* et qui, en même temps, ne contiennent pas le terme *physique*.

En théorie, un usage correct des opérateurs booléens devrait permettre de dépasser les problèmes mis en évidence pour l'indexation des pages Web. Jusqu'à il y a quelques années, il était possible, pour les navigateurs, de choisir si utiliser ou non les opérateurs booléens pour leurs interrogations, en accédant à la recherche avancée ou *advanced reserarch*. En outre, chaque moteur tendait avoir une syntaxe spécifique, en utilisant par exemple les caractères + ou & pour indiquer l'opérateur **AND**. Normalement, la liste et les fonctions de ces symboles se trouvaient dans la page *Help* ou Aide, qui pouvait être consultée à partir de la *Home Page* du moteur, et qui avait aussi la tâche d'indiquer les réglages de défaut (couramment, les premiers résultats étaient générés par **AND**, les suivants par **OR**). Dans des cas différents, à l'aide de la section *Advanced* ou recherche avancée, on pouvait aussi accéder directement aux expressions logiques précédemment définies. Dans cette dernière section, était importante la recherche par *exact phrase*, i.e. du syntagme exact, qui effectuait la recherche d'une séquence spécifique de mots.

Cependant, comme s'ils voulaient souligner l'insuffisance de leur sensibilité linguistique, les moteurs de recherche ont récemment éliminé les options explicites relatives à l'usage des opérateurs booléens, y compris celles de l'*exact phrase* et de l'*advanced search*. En fait, toute interrogation effectuée actuellement utilise, simultanément et par défaut, tous les opérateurs booléens, ce qui justifie le nombre élevé de résultats obtenus. Il s'agit d'un choix qui pénalise les usagers des moteurs du Web, et qui rend encore plus complexe la récupération d'informations sur le Web, et, en même temps, méconnaît l'utilité des ces opérateurs. Il est difficile de trouver des raisons valables pour justifier ce choix, même si ces applications booléennes n'étaient pas vraiment élémentaires ou même étaient peu fréquentes: leur mise de côté a représenté un désavantage pour les navigateurs plutôt qu'une aide.

La seule solution à ces problèmes pour les recherches sur le Web devient alors strictement dépendante de la réalisation d'index exhaustifs des pages présentes sur le réseau Internet. La recherche à l'aide des moteurs Web serait ainsi efficace et économique, si les indexations étaient effectuées sur la base du contenu complet des

textes, surtout en n'oubliant pas les différences formelles et sémantiques existant entre des mots simples comme *carte* et des mots composés comme *carte de crédit*; des différences qui, au contraire, sont prises en considération par les opérations de création et de structuration des dictionnaires électroniques que nous allons décrire. Il s'ensuit donc la nécessité de construire des dictionnaires et des index textuels complets et différenciés par unités de signification, i.e. pour un même texte, un dictionnaire de mots simples et un dictionnaire de mots composés. De cette façon, la recherche sur le Web peut être remarquablement plus précise et en même temps rapide, vu la puissance d'élaboration des plates-formes hardware et des logiciels les plus récents.

Un autre aspect qui ne semble pas être considéré comme il le devrait dans la création des index des pages Web, et par conséquent dans la recherche à l'aide des moteurs du réseau Internet, c'est l'ambiguïté du langage naturel, phénomène pour lequel un mot ou un énoncé peuvent avoir plus d'un signifié. Prenons par exemple les phrases italiennes suivantes:

1) *L'ufficio rise alla battuta di Max*

("Le bureau rit à la blague de Max")

2) *La vecchia curva la sbarra*

("La vieille femme courbe la barre", mais aussi "La vieille courbe la barre")

3) *Ieri ho mangiato un uovo in camicia*

("Hier j'ai mangé un œuf poché")

Ces phrases nous fournissent trois exemples différents d'ambiguïté, dont l'un est résolu grâce au contexte tandis que les autres nécessitent une segmentation spécifique ou une analyse plus approfondie des règles de co-occurrences et de restrictions de sélection.

En (1), l'ambiguïté est surtout lexicale et est liée au mot *ufficio* (bureau), dont nous donnons ici la définition extraite de *Lo Zingarelli 2002 in CD-Rom*:

*ufficio*²⁰ o (pop.) *ufficio spec. nel sign. 2, †ufficio, †oficio, †ofizio, (lett.) ufficio spec. nel sign. 9, †uficio, (raro) ufizio* [vc. dotta, lat. *officiu(m)*, che sta per *opificiu(m)*, da *opifex*, comp. di opus 'lavoro' e della radice di *facere* 'fare, compiere'; 1306]

²⁰ *ufficio* ou (pop.) *ufficio spéc. dans le sign. 2, †ufficio, †oficio, †ofizio, (lett.) ufficio spéc. dans le sign. 9, †uficio, (rare) ufizio* [voix docte, lat. *officiu(m)*, qui est utilisé au lieu de *opificiu(m)*, de *opifex*, comp. de opus 'lavoro' et de la racine de *facere* 'faire, accomplir'; 1306]

s. m.

1 (dés.) Ce que chacun doit faire selon le lieu, le temps, la condition, l'attitude, la préparation spécifique et sim.: **manquer à son propre office; à lui est l'office d'assister, de soigner; accomplir un office charitable envers les morts; office de mère, de tuteur, d'instituteur.** SIN: **devoir, obligation.**

2 (ext., litt.) **Bénéfice, faveur, service: office envers le prochain, envers soi même.**

3 (ext.) Intervention, recommandation, sollicitation: **ce que nous avons obtenu nous le devons à vos bons offices; il a interposé ses bons offices auprès du ministère. Bons offices, dans le droit international, médiation: requérir les bons offices d'un pays impartial.**

4 (ext.) Charge, tâche: **accepter, refuser un office; office épineux, délicat; office de témoin, arbitre, conciliateur.**

5 (droit) Ensemble de fonctions dont un fonctionnaire est chargé: **devoirs et charges de l'office. Office privé**, explication au nom propre d'une activité dans l'intérêt d'autrui, en obéissance à un commandement légal. **D'office**, par l'initiative autonome d'un fonctionnaire ou assimilé, sans une instance préalable: **démissions d'office; délit à poursuivre d'office. Acte d'office**, promulgué par un fonctionnaire ou une autre autorité dans l'exercice de ses fonctions. **Défenseur d'office**, désigné par le juge et assigné à la partie qui ne peut pas s'en procurer un de confiance. **Défense d'office**, celle faite par un défenseur d'office; (ext., fig.) argumentation de laquelle on n'est pas profondément convaincu, qu'on doit alléguer parce qu'obligé par d'autres personnes ou par des circonstances. (est.) Charge: **conférer l'office de Ministre.** (ext.) Lieu où un fonctionnaire exerce les fonctions qu'ils lui reviennent: **se rendre au cabinet du juge, du ministre.**

6 (organisation de l'entreprise) Tâche qu'une personne accomplit ou doit accomplir dans le cadre de l'organisation d'une entreprise: **négliger sa tâche; une tâche pleine de responsabilités.** (ext.) Lieu de travail d'un employé ou d'un dirigeant: **se rendre au bureau, aller, être au bureau.**

7 (droit) Organe: **bureau de placement, de renseignements. Bureau du siège**, autorité judiciaire ayant la fonction de juger. **Bureau de présidence**, ensemble du président et du vice-président d'un organe collégial. Le siège où ce même organe exerce ses fonctions. **Bureau électoral**, dans lequel on accomplit les opérations de vote ou de dépouillement de vote ou de calcul des résultats d'une élection.

8 Dans une entreprise publique ou privée, ensemble de fonctions d'entreprise homogènes, pour la plupart regroupées dans un seul secteur de la même entreprise, et aussi siège où elles sont accomplies: **bureau de vente; bureau d'expédition; bureau du personnel; bureau de propagande; bureau d'études.** (ext.) Ensemble d'employés qui accomplissent une activité spécifique soit dans l'entreprise,

s. **m. 1** (disus.) *Ciò che ciascuno deve fare secondo il luogo, il tempo, la condizione, l'attitudine, la preparazione specifica e sim.: mancare al proprio ufficio; a lui spetta l'ufficio di assistere, di curare; adempiere a un pietoso ufficio verso i defunti; ufficio di madre, di tutore, d'insegnante. SIN. Dovere, obbligo.*

2 (est., lett.) *Beneficio, favore, servizio: ufficio verso il prossimo, verso se stesso.*

3 (est.) *Intervento, raccomandazione, sollecitazione: ciò che abbiamo ottenuto lo dobbiamo ai vostri buoni uffici; ha interposto i suoi buoni uffici presso il ministro. Buoni uffici, nel diritto internazionale, mediazione: richiedere i buoni uffici di un Paese neutrale.*

4 (est.) *Incarico, incombenza: accettare, rifiutare un ufficio; ufficio spinoso, delicato; ufficio di padrino, arbitro, paciere.*

5 (dir.) *Insieme di funzioni di cui è investito un funzionario: doveri e oneri dell'ufficio. Ufficio privato, esplicazione in nome proprio di un'attività nell'interesse di altri, in ottemperanza di un comando legale. D'ufficio, per autonoma iniziativa di un funzionario, di un'autorità e sim., senza una previa istanza: dimissioni d'ufficio; reato perseguibile d'ufficio. Atto d'ufficio, emanato da un funzionario o da un'autorità nell'esercizio delle sue funzioni. Difensore d'ufficio, designato dal giudice e assegnato alla parte che non può procurarsene uno di fiducia Difesa d'ufficio, quella fatta da un difensore d'ufficio; (est., fig.) argomentazione di cui non si è intimamente convinti, che si deve addurre perché obbligati da altri o dalle circostanze. (est.) Carica: conferire l'ufficio di Ministro (est.) Luogo in cui un funzionario esercita le funzioni che gli competono: recarsi nell'ufficio del giudice, del ministro.*

6 (org. az.) *Compito che una persona svolge o deve svolgere nell'ambito dell'organizzazione di un'azienda: trascurare l'ufficio; un ufficio pieno di*

soit dans le lieu où ils travaillent: bureau de caisse; bureaux administratifs; être désigné chef de bureau: aller au bureau de poste, au bureau télégraphique.

9 *Prière, cérémonie, fonction religieuse: service funèbre. Office divin, liturgie avec laquelle l'église catholique sanctifie les différentes heures du jour, célébrées avec un chœur par des communautés religieuses ou singulièrement par les prêtres séculiers: réciter, dire l'office; chanter l'office. (ext.) Bréviaire, livre ou manuel qui contient les textes liturgiques à réciter pendant les différentes heures de l'office divin (c'est nous qui traduisons).*

responsabilità (est.) Posto di lavoro di un impiegato o di un dirigente: recarsi in ufficio; andare, essere in ufficio.

7 (dir.) Organo: ufficio di collocamento, d'informazioni Ufficio giudicante, autorità giudiziaria esplicante la funzione di giudicare Ufficio di presidenza, complesso del presidente e del vicepresidente di un organo collegiale La sede in cui lo stesso esplica le proprie funzioni Ufficio elettorale, nel quale si svolgono le operazioni di voto o di spoglio delle schede o di calcolo dei risultati di un'elezione.

8 In un'azienda pubblica o privata, complesso di funzioni aziendali omogenee, per lo più raggruppate in un unico settore della stessa, e sede in cui sono svolte: ufficio vendite; ufficio spedizioni; ufficio personale; ufficio stampa; ufficio propaganda; ufficio studi (est.) Complesso di impiegati che svolgono una determinata attività nell'ambito dell'azienda, e sede in cui lavorano: ufficio cassa; uffici amministrativi; essere nominato capo ufficio; recarsi all'ufficio postale, telegrafico.

9 Preghiera, cerimonia, funzione religiosa: ufficio funebre Ufficio divino, liturgia con la quale la Chiesa cattolica santifica le diverse ore del giorno, celebrata in coro da alcune comunità religiose o singolarmente dai preti secolari: recitare, dire l'ufficio; cantare l'ufficio (est.) Breviario, libro o manuale che contiene i testi liturgici da recitarsi nelle varie ore dell'ufficio divino.

Le mot *ufficio* est donc très ambigu, parce qu'il peut avoir jusqu'à neuf significés différents, selon les contextes dans lesquels il est utilisé. Ce type d'ambiguïté est du type lexical, et en ce qui concerne la phrase (1), elle est résolue par le rapport existant entre le verbe *ridere* (rire) et son sujet, c'est-à-dire *l'ufficio*, qui donne à ce mot le signifié *d'ensemble d'employés qui accomplissent une activité spécifique soit dans l'entreprise, soit dans le lieu où ils travaillent*²¹. En fait, le verbe *ridere* de la phrase (1) prédique une

²¹ Nous pouvons noter que ce signifié est inséré dans la note d'usage 8, dans laquelle nous trouvons aussi celle relative à *dans une entreprise publique ou privée, ensemble de fonctions d'entreprise homogènes*, qui appartient à un domaine sémantique différent de celui que nous sommes en train d'analyser. De ce point de vue, il aurait été plus correct de séparer les deux définitions, afin d'indiquer un usage spécifique pour *ensemble d'employés qui accomplissent une activité spécifique soit dans l'entreprise, soit dans le lieu où ils travaillent*.

action qui peut être accomplie seulement par des humains, au moins dans des phrases qui ont un sens communément acceptable et non surréaliste, et en ce cas le sujet du verbe, c'est-à-dire *l'ufficio* (Le bureau), comme nom singulier collectif, ne peut indiquer qu'un groupe de personnes en état de rire.

L'ambiguïté lexicale est un phénomène très courant et qui peut concerner encore d'autres mots, comme par exemple *civile* ("civil", comme nom et adjectif, et aussi "civique"), *corte* (le nom "cour", mais aussi "audience" ainsi que "courtes", adjectif féminin pluriel de "court"), *imputato* ("accusé" comme nom, mais aussi "accusé" et "imputé" comme participes passés et adjectifs), et *penale* ("pénal" et "judiciaire" comme adjectifs, et aussi "amende" et "pénalité" comme noms), qui ont des niveaux d'occurrence très hauts dans les recherches Web parce qu'ils sont utilisés dans les interrogations pour la récupération d'information concernant la jurisprudence. A propos de ces termes, nous pouvons ajouter:

- a) *civile*: c'est un substantif masculin et féminin singulier **mais aussi** un adjectif masculin et féminin singulier **et aussi** un substantif seulement singulier (le *civile* identifié comme domaine d'application du code civil);
- b) *corte*: c'est un substantif féminin singulier **mais aussi** un adjectif féminin pluriel (comme forme fléchie de *corto*);
- c) *imputato*: c'est un substantif masculin singulier **mais aussi** un adjectif masculin et féminin singulier **et aussi** le participe passé masculin singulier du verbe *imputare* (inculper, accuser);

- d) *penale*: c'est un substantif féminin singulier **mais aussi** un adjectif masculin et féminin singulier **et aussi** un substantif masculin seulement singulier (le *penale* identifié comme domaine d'application du code pénal).

En n'oubliant pas ce que nous avons dit à propos d'*ufficio*, les trois exemples précédents nous amènent à supposer qu'en choisissant par exemple *corte* comme terme à rechercher à l'aide d'un moteur Web, il sera possible de trouver des liens vers des pages dans lesquelles ce mot pourra avoir la fonction d'adjectif, outre celle de nom. Cette même observation est valable pour les exemples des points (a), (c) et (d), et démontre que les recherches Web effectuées seulement sur la base de l'aspect formel de mots simples, c'est-à-dire en ne considérant pas leurs propriétés morphologiques et grammaticales, nous donneront des résultats dont la crédibilité sera directement proportionnelle à l'ambiguïté lexicale de ces mêmes mots. Donc, si leur niveau d'ambiguïté est élevé, les mots utilisés dans l'interrogation ne seront pas capables de représenter l'information contenue dans les différentes pages Web. Un utilisateur humain pourra aisément résoudre les ambiguïtés sur des bases contextuelles, mais pour récupérer les informations requises il sera néanmoins obligé de lire toutes les pages, même celles qui ne concernent pas le sujet de la recherche.

Au contraire, la phrase (2) présente un type d'ambiguïté liée à l'ordre dans lequel les mots apparaissent et aux catégories grammaticales qu'il est possible de lui assigner, et nous l'appellerons donc *ambiguïté syntactico-grammaticale*. Cette phrase peut donc être étiquetée comme suit:

2a)	<i>La</i>	<i>vecchia</i>	<i>curva</i>	<i>la</i>	<i>sbarra</i>
2b)	<i>Dét</i>	<i>Adj</i>	<i>N</i>	<i>Pron</i>	<i>V</i>
2c)	<i>Dét</i>	<i>N</i>	<i>V</i>	<i>Dét</i>	<i>N</i>

En considérant que *Dét* indique un *déterminant* (c'est-à-dire un *article*), *Adj* un *adjectif*, *N* un *nom*, *V* un *verbe* et *Pron* un *pronom*, nous pouvons noter que si en (2b) nous interprétons et étiquetons *vecchia* comme adjectif, afin d'assigner un signifié logique à la phrase nous serons obligés d'étiqueter *curva la sbarra* comme une séquence du type nom-prénom-verbe. Au contraire, en (2c), en interprétant et étiquetant *vecchia* comme un nom, toujours pour donner à la phrase un signifié logique, nous serons obligés d'étiqueter *curva la sbarra* comme une séquence verbe-déterminant-nom. Ce type d'analyse, qui est dit à états finis et qui est l'une des principales fonctions inférencielles des *parsers* et des *parsing* déjà décrits (même s'il est rigide fondé sur les règles syntaxiques d'une langue, dans ce cas l'italien, et même s'il est applicable avec succès à d'autres langues), n'arrive pas à résoudre l'ambiguïté de notre phrase ni d'autres phrases similaires. Ce type d'analyse peut toutefois mettre en évidence cette ambiguïté et donner les moyens pour segmenter cette phrase, en indiquant toutes les interprétations logiques possibles et grammaticalement acceptables d'une phrase donnée.

L'ambiguïté de la phrase (3) est par contre liée aux différentes segmentations que nous pouvons effectuer sur les éléments qui la composent, surtout en ce qui concerne la séquence *Ieri ho mangiato un uovo in camicia* (qui peut signifier : Hier j'ai mangé un œuf poché – Hier j'ai mangé un œuf, étant en chemise) comme le montrent les tables qui suivent²²:

²² Traduite en français, cette phrase perd son ambiguïté, qui provient du fait que la séquence *in camicia* peut modifier soit *uovo* que le pronom personnel *Io* (effacé), sujet du verbe *ho mangiato*.

3a)	<i>ieri</i>	<i>ho mangiato</i>	<i>un</i>	<i>uovo in camicia</i>
3b)	<i>Adv</i>	<i>V</i>	<i>Dét</i>	<i>NPN</i>

3c)	<i>ieri</i>	<i>ho mangiato</i>	<i>un</i>	<i>uovo</i>	<i>in camicia</i>
3d)	<i>Adv</i>	<i>V</i>	<i>Dét</i>	<i>N</i>	<i>Adv (PN)</i>

Les nouvelles étiquettes insérées dans ces descriptions indiquent que *Adv* désigne un *adverbe*, *NPN* un mot composé formé par une séquence du type nom-préposition-nom et *Adv (PN)* un adverbe composé avec une séquence du type préposition-nom. Si en (3a) la séquence *uovo in camicia* est un complément de *mangiare*, en (3c) la séquence *in camicia* est séparée de *uovo*, qui reste le seul complément de *mangiare*, et la phrase peut donc être réécrite comme suit:

3e) *Ieri, in camicia, ho mangiato un uovo*

(Hier, en chemise, j'ai mangé un oeuf)

Dans (3e), *ieri* et *in camicia* peuvent changer de place, et cette possibilité confirme que *in camicia* peut aussi être étiqueté et classifié comme un adverbe ayant la même distribution que *ieri*:

3f) *In camicia, ieri, ho mangiato un uovo*

(En chemise, hier, j'ai mangé un oeuf)

Les types d'ambiguïté mis en évidence, spécifiques du langage naturel, relèvent de la créativité des locuteurs, ne sont pas prévisibles ni ne peuvent être éludés et sont souvent présents dans les textes, y compris ceux des pages Web. Comme nous l'avons déjà dit, si nous utilisons un mot ambigu pour une interrogation, le moteur de recherche nous donnera des résultats aussi ambigus que le mot, c'est-à-dire nous amènera à consulter des pages Internet dans lesquelles ce mot sera localisé seulement comme séquence spécifique de lettres, et non sur la base de ses caractéristiques morphologiques, grammaticales et syntaxiques. Pour les moteurs de recherche, l'impossibilité d'ajouter une exploration linguistique à celle formelle déjà existante, pourrait être dépassée par un moteur linguistique ou un dictionnaire électronique étiqueté, où à chaque mot serait assignée une étiquette grammaticale univoque et où les mots appartenant à plus d'une catégorie grammaticale seraient lemmatisés séparément. Un tel dictionnaire, exhaustif du point de vue descriptif, nous permettrait même de résoudre des ambiguïtés plus complexes, comme celles mises en évidence en (2), vu qu'il donnerait aux logiciels de récupération des informations et aussi tous les éléments utiles pour effectuer le parsing minimal d'un texte. Comme nous le verrons, ces caractéristiques seront insérées dans le dictionnaire électronique dont nous donnerons la construction, et elles seront aussi utilisées à l'intérieur d'un logiciel d'analyse textuelle automatique qui, grâce à ses qualités spécifiques, pourrait être facilement utilisé avec succès pour la récupération des informations sur le Web. De façon schématique, nous résumons les particularités d'un dictionnaire électronique qui doit, à notre avis, être structuré de façon à comporter:

- la classification taxinomique du lexique, c'est-à-dire l'insertion dans la liste des lemmes de tout terme utilisé dans n'importe quel contexte médiatique et de divulgation, y compris ceux techniques et technologiques;
- une liste de lemmes continuellement mise à jour, de façon à ce qu'elle puisse inclure tous les néologismes d'une langue, sans, par contre, effacer les termes considérés comme obsolètes;

- des informations morphologiques et grammaticales explicites pour chaque lemme, qui incluent des éléments formels utiles pour résoudre l'ambiguïté entre homographes ou mots polysémiques. De cette façon, dans le dictionnaire électronique, les unités lexicales formellement similaires mais ayant des caractéristiques morphologiques et grammaticales différentes, deviennent des entrées distinctes et sont lemmatisées séparément. Ceci fait que dans le dictionnaire électronique il n'y a pas d'entrées doubles.

1.3.2 La correction privée des *spelling checkers*

Comme nous l'avons déjà remarqué, l'un des avantages de l'écriture informatisée est la possibilité de réviser et de corriger automatiquement les textes réalisés ou importés sur ordinateur. Ces types de correction peuvent concerner les coquilles, l'observation des règles de grammaire – non du type analysé et prédit avec le parsing – et aussi la consultation d'un Thesaurus, avec lequel il est possible de vérifier et substituer synonymes et contraires.

La création de ce type de logiciels, dits en anglais *spelling and grammar checkers* et souvent intégrés aux word processors, a représenté un élément très important dans le domaine du langage bureautique et commercial, parce qu'il semblait pouvoir accélérer des opérations notamment longues et laborieuses, faites principalement "à la main". Néanmoins, avec le temps, pour ce type de logiciels comme pour les moteurs de recherche Web, on s'est aperçu que les parties linguistiques nécessitaient une approche méthodologique différente, surtout en ce qui concernait la structuration du moteur linguistique avec lequel il faut effectuer les révisions. Aujourd'hui, cette nouvelle approche n'est pas encore à l'horizon ni elle n'est prospectable, et les carences que nous allons analyser pour ces logiciels ont, comme pour les moteurs de recherche, de sérieuses rechutes pour leurs fonctionnalités et finalités.

Notre analyse concernera le contrôle de l'orthographe et l'utilisation du Thesaurus, mais nous ne nous occuperons pas de la correction grammaticale, qui ne s'appuie pas seulement sur un moteur linguistique ou un dictionnaire électronique. Cependant, par rapport à la correction grammaticale automatique, nous pouvons affirmer qu'en général elle est effectuée d'une façon plutôt bizarre, parfois en appliquant des règles qui, dans la meilleure des hypothèses sont obsolètes ou abstruses, donc visiblement restrictives de la créativité des locuteurs. Par conséquent, il est envisageable qu'après quelques tentatives, pour un usager ordinaire la correction grammaticale à l'aide d'un word processor puisse revêtir une valeur presque nulle.

La révision orthographique d'un *spelling checker* prévoit que les mots d'un texte écrits par l'ordinateur sont formellement confrontés à ceux insérés dans un dictionnaire électronique intégré dans le même logiciel. Le contrôle peut être laissé toujours actif en background, et de cette façon le word processor, pendant la frappe souligne en rouge ou met en évidence en d'autres façons les mots non reconnus, soit qu'il s'agit de coquilles soit de termes non présents dans le dictionnaire intégré. Ceci contient néanmoins seulement une partie et non toutes les entrées lexicales d'une langue donnée, et pour cette raison peut être mis à jour directement par l'utilisateur.

Initialement, les mots contenus par un dictionnaire électronique intégré dans un *spelling checker* ne sont donc pas nombreux, et pour cette raison il peut arriver souvent que pendant la frappe ou la révision orthographique d'un document, le word processor indique comme inconnus des termes plutôt communs en italien, tels que par exemple *aggiornabilità* (possibilité de mettre ou être mis à jour), *epistemologicamente* (épistémologiquement), *burotico* (burotique) et *inferenziale* (inférentiel) ou formellement singuliers – c'est-à-dire quelques enclitiques particuliers²³ tel que *esprimentesi* (qui est en train de s'exprimer) et *dicansi* (à dire) ou les polyrhémétiques du type *altomedievale* (de la haute Moyen Age) et *nordoccidentale* (nord-occidental).

Il est important de ne pas oublier que ceux qui écrivent ou lisent un texte à l'ordinateur sont obligés de s'arrêter à chaque fois qu'ils rencontrent un élément quelconque mis en évidence par le *spelling checker*, pour contrôler s'il s'agit d'une coquille ou d'un terme inconnu. Seulement tout cela serait déjà suffisant pour ralentir les opérations de révision, mais celles-ci décélèrent ultérieurement si le moteur linguistique du *spelling checker* n'est pas particulièrement riche en entrées, parce qu'il mettra en évidence dans le texte beaucoup de termes écrits correctement, pour le simple fait qu'il ne les reconnaîtra pas, contraignant ainsi le lecteur/écrivain à faire des arrêts nombreux et inutiles.

En même temps, la possibilité de mettre à jour le dictionnaire d'un *spelling checker* représente une source ultérieure de ralentissement pour les opérations de révision, parce qu'elle met les usagers devant des doutes de nature linguistique. Souvent,

²³ Voir Monteleone, M. 1996.

par rapport aux termes sous contrôle, une fois vérifié que leur orthographe est correcte, on se demande par exemple s'ils sont réellement en usage, donc compréhensibles aussi pour d'autres locuteurs de la même langue. Dans ces cas, il est presque obligatoire de recourir à des oeuvres classiques de consultation telles que les dictionnaires papier, mais cette opération ralentit irrémédiablement les démarches de révision automatique, parfois en ne dissipant même pas les doutes qui se sont présentés.

Donc, des soi-disant instruments essentiels pour la structure portante d'un *spelling checker* en déterminent au contraire la faillite, qui pourrait être évitée si le moteur linguistique d'un pareil logiciel avait parmi ses buts celui de cataloguer taxinomiquement le lexique d'une langue, c'est-à-dire s'il incluait tous les mots en usage de celle-ci.

Il faut dire aussi que la vérification exclusivement formelle d'un texte ne permet pas d'affronter correctement un certain type d'ambiguïtés qui dérivent de l'usage des homographes, c'est-à-dire de ces mots qui ont des formes identiques mais qui appartiennent à des catégories grammaticales différentes. En ce sens, un type très particulier de faute provient directement de la structuration du *Thesaurus* – ou dictionnaire des synonymes et contraires – intégré dans les différentes versions de Microsoft Word®. Si par exemple dans un document ouvert par ce word processor nous écrivons le mot italien:

eluse (*éluda* mais aussi *éludées*)

qui d'un point de vue morphologique et grammatical peut être soit l'indicatif passé simple du verbe *eludere* (éluder), soit le participe passé féminin pluriel du même verbe, et si nous cherchons un synonyme de ce mot dans le *Thesaurus* de Microsoft Word®, nous obtenons la liste suivante:

- *evitò* (évita)

- *ingannò* (trompa)

qui contient seulement des verbes et aucun synonyme pour le participe passé, même si dans cette catégorie grammaticale ce mot en a plusieurs, comme:

- *evitato* (évit  ), *schivato* (esquiv  ), *scansato* (esquiv  )
- *ingannato* (tromp  ), *raggirato* (circonvenu), *trasgredito* (transgress  )

Ces synonymes sont correctement mentionn  s dans d'autres dictionnaires italiens, sur papier ou informatiss  s, et il n'y a aucune raison plausible pour laquelle le *Thesaurus* de Microsoft Word   doit constituer une exception inexplicable.

1.4 LA “SEPARATION DOLOUREUSE”

Nous avons vu que les dictionnaires papier et les dictionnaires électroniques ont des aspects et des caractéristiques communes, en ce qui concerne soit leur structure soit leurs finalités descriptives et applicatives. Dorénavant, nous prendrons en considération les différences qui existent entre ces deux types de dictionnaire, qui d'un côté délimiteront les usages et les fonctionnalités du dictionnaire électronique, et de l'autre mettront en évidence quelques limites des dictionnaires papier.

Comme beaucoup d'autres activités, la presse a été informatisée et donc modernisée, et en quelques cas la commercialisation des dictionnaires papier est aujourd'hui faite sur supports informatiques. De toute façon, la formalisation des caractères et des règles typographiques n'a pas influencé ni apporté des changements au contenu des publications. Du point de vue du dictionnaire papier, l'informatisation a donc seulement introduit de nouveaux supports textuels, de nouvelles techniques typographiques et de nouveaux procédés de consultation. Informatiser un dictionnaire papier, n'équivaut pas à réaliser un dictionnaire électronique, même si l'ouvrage sur papier peut de toute façon être l'une des sources principales pour la création de dictionnaire électronique.

Maurice Gross²⁴ individualise les canons précis suivant lesquels il est possible d'obtenir un dictionnaire électronique en en informatisant un sur papier, et à cette fin il théorise une procédure d'*information retrieval*, pendant laquelle la partie textuelle des dictionnaires papier est lue et explorée par des logiciels spécifiques, afin d'en extraire automatiquement toutes ces expressions qui listent toujours dans la même forme et qui peuvent donc être utilisées pour la construction d'une base de données lexicale homogène. Cette procédure semble être possible parce que les parties textuelles des dictionnaires papier sont brèves, simples, souvent stéréotypées et donc susceptibles d'une analyse automatique plus efficace par rapport aux parties textuelles plus longues. On pourrait donc théoriser que dans les dictionnaires papier la récupération des informations à l'aide de mots-clefs, très semblable à celle utilisée pour les moteurs de

²⁴ Voir AA.VV., 1989.

recherche du Web, pourrait avoir une bonne réussite, surtout si on étendait la recherche en utilisant aussi des synonymes.

Néanmoins, il existe un obstacle important au succès de cette opération, c'est le fait que les informations données dans un dictionnaire papier pour les lemmes – ou pour les groupes de lemmes sémantiquement contigus entre eux – n'ont ni une description ni une forme identique. Pour mettre en évidence cet aspect, nous analyserons quelques noms italiens de profession, comme ils sont décrits dans *Lo Zingarelli 2001 in CD-Rom*, en traduisant une partie des descriptions de chaque entrée:

1. **brigadiere s. m.** *Sottufficiale dell'Arma dei Carabinieri e del Corpo della Guardia di Finanza, che ha grado corrispondente a quello di sergente maggiore delle altre Armi.*
2. **carabiniere s. m.** *1) Un tempo, soldato a piedi o a cavallo, armato di carabina. 2 f. -a) Appartenente all'Arma dei Carabinieri che svolge compiti di polizia civile, militare e giudiziaria: carabiniere a piedi, a cavallo / **Fare il carabiniere, essere un carabiniere**, (fig.) comportarsi in modo particolarmente severo e autoritario.*
3. **finanziere s. m. (f. -a)** *1) Chi tratta affari di alta finanza / Chi si occupa di problemi finanziari. 2) Membro del corpo delle guardie di finanza.*
4. **sottotenente s. m.** *Primo grado della gerarchia degli ufficiali.*

5. **tenente s. m.** *Deuxième grade de la hiérarchie des officiers, suivant celui de sous-lieutenant, auquel est confié le commandement d'un peloton des armes différentes ou de la ligne des pièces d'une batterie d'artillerie.*²⁵

Même si les unités lexicales appartiennent au même domaine sémantique, dans les quatre cas précédents nous notons l'application de modalités descriptives extrêmement variables. En fait, les descriptions de (1), (3) et (4) sont structurées de façon différente de celle de (2), qui est introduite par le pronom *qui*, tandis qu'en termes morphologiques seulement pour (2) et (3) on a indiqué aussi les flexions au féminin, dont la notation serait toutefois justifiée pour les autres lemmes aussi. En fait, *tenente* et son composé *sottotenente*, sont des mots masculins ou féminins, aussi bien au singulier qu'au pluriel. En outre, l'indication de la forme féminine possible pour *finanziere*, même si elle n'est pas assez commune, justifierait une notation identique pour le lemme *brigadiere*. Comme dernière considération, nous pouvons aussi observer que (2) et (3) sont les seuls lemmes pour lesquels sont indiqués deux usages possibles, en vertu desquels il aurait été peut-être plus logique d'effectuer des lemmatisations séparées.

Par rapport au domaine des carrières militaires, si nous voulions extraire automatiquement du *Zingarelli 2001 in Cd-Rom* tous les grades possibles, nous devons présupposer qu'à l'intérieur de ce dictionnaire:

-
1. **brigadier** s. m. Sous-officier de l'Arma dei Carabinieri et du Corpo della Guardia di Finanza, qui a un grade correspondant à celui de sergent-major des autres Armes.
 2. **carabinier** s. m. 1) Autrefois, soldat à pieds ou à cheval, armé de carabine. 2 (f. -a) Individu qui appartient à l'Arma dei Carabinieri et qui a des tâches de surveillance civile, militaire et judiciaire: carabinier à pieds, à cheval. Faire le carabinier, être un carabinier, (fig.) se porter d'une façon particulièrement sévère et autoritaire.
 3. **financier** s. m. (f. -a) 1) Qui traite d'affaires de haute finance. Qui s'occupe de problèmes financiers. 2) Membre du corps des Guardie di Finanza.
 4. **sous-lieutenant** s. m. Premier grade de la hiérarchie des officiers.
 5. **lieutenant** s. m. Deuxième grade de la hiérarchie des officiers, suivant celui de sous-lieutenant, auquel est confié le commandement d'un peloton des armes différentes ou de la ligne des pièces d'une batterie d'artillerie.

(C'est nous qui traduisons). Il faut souligner que dans *Lo Zingarelli 2004 in CD-Rom*, on a modifié la classification morphologique de 4) et 5) en "s. m. et f." (substantif masculin et féminin).

- les descriptions sémantiques pour ces noms sont formellement toutes identiques, c'est-à-dire qu'elles utilisent toujours les mêmes éléments linguistiques. Cela signifierait qu'une séquence distinctive, par exemple *grade de la hiérarchie de...* devrait apparaître seulement dans les gloses relatives aux grades militaires, comme si elle était une étiquette identificatrice. Comme nous l'avons déjà vu, une telle systématique descriptive est une propriété essentielle pour la réalisation de systèmes automatiques d'*information retrieval*;
- la description morphologique et grammaticale de ces mots doit être homogène, c'est-à-dire qu'elle soit appliquée uniformément à tous les termes qui appartiennent à ce domaine sémantique spécifique. Dans notre cas, cela signifie que pour tous les noms relatifs aux grades de la carrière militaire il doit toujours être possible de trouver l'indication relative à la flexion féminine singulière et plurielle, ce qui permettrait de faire des distinctions morphologiques spécifiques entre les homographes d'une même classe grammaticale – par exemple, entre le *carabiniere* avec le sens de *soldat à pieds ou à cheval, armé de carabine* et *carabiniere* avec le sens d'*individu qui appartient à l'Arma dei Carabinieri* (féminin en *carabiniera*).

Les exemples de (1) à (5) démontrent au contraire que les deux conditions précédentes ne sont pas observables dans *Lo Zingarelli 2001*, et que donc à partir des descriptions morphologiques et grammaticales de ce dictionnaire il ne serait pas possible d'extraire automatiquement les informations relatives à l'ensemble sémantique des grades des carrières militaires. L'absence de ces deux conditions est d'ailleurs commune à beaucoup de dictionnaires papiers et d'encyclopédies aujourd'hui dans le commerce. Elle porte directement à conclure que pendant la structuration d'un dictionnaire électronique, dans une première phase de repérage des unités lexicales, la simple conversion informatique en base de données lexicales des données sur papier non seulement ne pourra se servir de la consultation automatique des sources sur papier, mais ne sera même pas suffisante pour créer une base de données lexicales homogène. Cette considération est aussi valable pour le contenu des éditions sur CD-ROM, dans

lesquels les descriptions ne présentent pas de modifications par rapport aux versions respectives sur papier. Par conséquent, pour la construction d'un dictionnaire électronique la seule partie qu'on pourra transposer à partir des dictionnaires papier sera la liste de lemmes, et en outre cette transposition devra être faite par *data entry*, et ne pourra donc être assez rapide, surtout si l'on considère qu'un dictionnaire papier est constitué d'au moins 80.000 lemmes.

Les différences les plus importantes entre dictionnaires papier et dictionnaires électroniques sont évidentes. Même si les deux sont consacrés à un seul objectif, la description du lexique d'une langue, les premiers reflètent la pluralité expressive du langage naturel, tandis que les seconds ont la nécessité de cataloguer les éléments lexicaux, d'en individualiser les aspects formels et de les décrire d'une façon homogène. Les uns n'ont pas d'intérêt pour l'uniformité descriptive, les autres font de cette caractéristique une prérogative essentielle. Il s'agit de différences méthodologiques très importantes qui néanmoins n'ont aucune raison d'être affadies: en fait, il n'y a aucune motivation raisonnable – commerciale ou scientifique – qui puisse pousser les éditeurs à revoir, dans les termes que nous venons d'indiquer, le contenu de leurs ouvrages sur papier. La vente et le succès des dictionnaires papier ne dépendent pas de l'exposition formellement homogène du contenu, mais de l'idéologie descriptive au rapport qualité prix, en passant par la présentation esthétique et la maniabilité. Il est même à supposer que si un dictionnaire papier utilisait une série de formules toujours égales pour décrire le lexique, pourrait se révéler monotone à la consultation et même peu didactique, bref repoussant. Décrire un lemme avec d'autres mots - sémantiquement, morphologiquement ou grammaticalement - est une opération complexe qui doit pouvoir profiter de toutes les potentialités expressives d'une langue. Mais pour les raisons que nous avons analysées, le dictionnaire électronique, soit comme base de données soit comme instrument linguistique-informatique, ne peut pas faire abstraction d'une description formellement homogène de son contenu; cette nécessité complète la phase de séparation par le modèle sur papier et justifie l'existence des différences formelles et méthodologiques entre les deux types de dictionnaires.

Par contre, il faut faire une autre considération pour ce qui concerne les différences descriptives du lexique, qui comme nous le verrons intéressent la vision

morphologique, grammaticale et sociolinguistique d'une langue donnée. En ce sens, et en revenant sur les canons indiqués pour la création d'une base de données lexicale, l'observation des situations réelles et l'efficacité de leur description sont des aspects non seulement souhaitables, mais nécessaires pour le catalogage correct du lexique, sans lequel un usage cohérent du dictionnaire électronique comme instrument linguistique-informatique n'est pas possible. En ce domaine les différences entre les deux types de dictionnaires prennent des proportions importantes, au désavantage de l'ouvrage sur papier.

CHAPITRE II. ANATOMIE D'UN DICTIONNAIRE ELECTRONIQUE

Dans ce chapitre, nous nous occuperons des modalités de réalisation d'un dictionnaire électronique, en analysant les bases lexicales sur lesquelles se fonde cette opération et aussi les procédés informatiques et de compilation nécessaires à sa création. Jusqu'à présent, nous avons défini le dictionnaire électronique surtout par opposition au dictionnaire papier; maintenant nous le définirons à l'intérieur de son domaine applicatif effectif, c'est-à-dire la linguistique-informatique, et en relation avec les théories du lexique-grammaire²⁶ dont il est l'un des instruments principaux. Nous verrons aussi que le dictionnaire électronique a un rôle de véritable moteur linguistique à l'intérieur des routines informatiques, en particulier d'Intex®²⁷, un paquet de logiciels créés sur la base des théories lexico-grammaticales conçues en France par Maurice Gross et développées en Italie par Annibale Elia.

Les dictionnaires électroniques que nous décrirons ici sont principalement de deux types, et sont subdivisibles sur la base de l'aspect formel et sémantique de leur contenu. Nous aurons donc:

- des dictionnaires électroniques de mots simples, qui incluent tous les mots sémantiquement autonomes et composés par des séquences de lettres sans interruption, comme *casa* (maison) ou *battello* (bateau);
- des dictionnaires électroniques de mots composés, qui incluent toutes les séquences formées par deux ou plus de deux mots qui composent conjointement des unités simples de signifié, telles que par exemple *casa di cura* (maison de santé) ou *battello a vapore* (bateau à vapeur).

²⁶ Voir Gross, M., 1985.

²⁷ Voir Silberztein, M., 1993.

Cette distinction est fondamentale parce que les deux types d'entrées que nous venons d'illustrer sont sémantiquement et formellement différentes et ont des caractéristiques morphologiques spécifiques. De ce point de vue, nous verrons par exemple que les modalités de flexion automatique des mots simples et des mots composés ne peuvent être effectuées simultanément, tandis qu'en termes de signifié et d'usage les mots composés se montreront moins ambigus que les simples. D'un point de vue informatique, la distinction formelle entre les deux types d'entrées est cruciale en vue du traitement automatique des dictionnaires et aussi de l'analyse textuelle automatique pour laquelle ces dictionnaires auront fonction de moteurs linguistiques.

Nous verrons comment naissent et sont structurés ces deux types de bases de données.

2.1 LE LEXIQUE ETIQUETE

Un dictionnaire électronique de mots simples se compose de quatre parties essentielles:

- 1) une liste de lemmes;
- 2) des codes alphanumériques;
- 3) des algorithmes pour la flexion automatique des lemmes, associés de façon univoque aux code alphanumériques;
- 4) des routines pour la flexion automatique, élaborées sur la base des algorithmes de flexion.

Ces parties ne sont toutefois pas indépendantes, mais sont en stricte relation même si elles prévoient des phases séparées de planification.

La transcription de la liste de lemmes sera le premier pas vers la création d'un dictionnaire électronique et elle comportera une phase préliminaire importante, c'est-à-dire l'individuation et la consultation de sources adéquates au repérage des données à insérer. En général, les sources sont non seulement les dictionnaires papier, mais aussi tous les autres contextes médiatiques de divulgation qui présentent une crédibilité sociolinguistique suffisante, c'est-à-dire tous ces domaines communicatifs dans lesquels on utilise la langue standard d'une façon plus ou moins déclarée. Néanmoins, cet élément minimal de discrimination n'est pas suffisant pour nous assurer que notre liste de lemmes, une fois complète, contiendra tous les mots utilisés ou utilisables dans une langue donnée, et il sera nécessaire d'appliquer d'autres normes et d'effectuer des tests spécifiques pour définir quels termes devront être insérés dans la liste de lemmes d'un dictionnaire électronique. En outre, il ne faudra pas compléter l'analyse de toutes les sources possibles avant de passer à la phase pratique de création de la liste de lemmes,

qui comportera un long travail de *data entry*. La quantité de sources analysables se présente potentiellement comme très vaste, et il sera donc convenable de se focaliser sur la création d'une liste de lemmes de base²⁸, qui puisse être mise à jour pendant des phases successives mais qui inclut les termes plus fréquemment utilisés dans une langue.

En commençant la phase de lemmatisation, il sera important d'effectuer une distinction préliminaire entre mots fléchis et non fléchis. Il s'agit d'une différenciation formelle très importante, parce que d'un point de vue morphologique cela nous aidera à définir deux sous-ensembles spécifiques du lexique d'une langue, avec des caractéristiques distinctives qui seront en conséquence formalisées par des algorithmes de flexion également distinctifs. Donc, les entrées flexionnelles de la liste de lemmes seront insérées comme suit:

- les verbes à l'infinitif *amare* (aimer), *vivere* (vivre), *tornare* (revenir);
- tous les mots épicènes au singulier *guardia* (garde), *pantera* (panthère), *sosia* (sosie);
- tous les mots non épicènes au masculin singulier *amministratore* (administrateur), *deputato* (député), *insegnante* (enseignant).

Les entrées non flexionnelles auront la lemmatisation suivante:

²⁸ Pour la définition de lexique de base de l'italien, voir De Mauro, T., 1990.

- toutes sous leur forme invariable *apparentemente* (apparemment), *domani* (demain), *sopra* (sous, comme adverbe et comme préposition).

Chaque entrée sera accompagnée d'une étiquette qui en indiquera la catégorie grammaticale d'appartenance. Les étiquettes assignables seront les suivantes:

- **A** pour les adjectifs
- **AVV** pour les adverbes
- **CONG** pour les conjonctions
- **DET** pour les déterminants
- **ESC** pour les interjections
- **N** pour les noms
- **PAA** pour les prépositions composées formées par un déterminant et la préposition simple *a* (à), comme *al* (au)
- **PAC** pour les prépositions composées formées par un déterminant et la préposition simple *con* (avec), comme *col* (avec le)
- **PADA** pour les prépositions composées formées par un déterminant et la préposition simple *da* (de), comme *dal* (du)
- **PADI** pour les prépositions composées formées par un déterminant et la préposition simple *di* (de), comme *del* (du)
- **PAN** pour les prépositions composées formées par un déterminant et la préposition simple *in* (in), comme *nel* (dans le);
- **PAS** pour les prépositions composées formées par un déterminant et la préposition simple *su* (sous), comme *sul* (sous le)

- **PREP** pour les prépositions simples
- **PRON** pour les pronoms
- **V** pour les verbes

Si une entrée appartient à plus d'une catégorie grammaticale, dans la liste des lemmes elle est redoublée, c'est-à-dire que pour cette entrée on crée deux chaînes avec des étiquettes grammaticales différentes, comme pour le mot *rosso* (rouge), qui peut être un nom et aussi un adjectif:

rosso,.A

rosso,.N

Simultanément à l'assignation des étiquettes, on créera et on assignera aux entrées les codes numériques de flexion, dont la réalisation se basera sur une étude empirique des mutations morphologiques des entrées flexionnelles, en relation avec les propriétés de genre (masculin et féminin) et de nombre (singulier et pluriel). Donc, en prenant comme exemple un mot tel que:

casa (maison)

on en repérera en premier lieu le genre, et ensuite toutes les formes fléchies possibles. En considérant que le mot de notre exemple est féminin et a une seule forme fléchie au pluriel, c'est-à-dire:

case (maisons)

pour en établir le modèle flexionnel dans toutes les formes, on séparera la partie commune (en ce cas la racine *cas-*) des parties qui changent (les désinences flexionnelles *-a* et *-e*), en créant ce schéma:

ms	fs	mp	fp
-	a	-	e

Table 1

Dans la table 1 les notations de la première ligne donnent les informations de genre et de nombre, sous forme d'étiquettes (**m**asculin **s**ingulier, **f**éminin **s**ingulier, **m**asculin **p**luriel, **f**éminin **p**luriel). Pour le mot féminin *casa*, les cases du genre masculin sont vides. Au contraire, dans la seconde ligne, sont insérées les désinences flexionnelles à ajouter à la racine pour obtenir les formes flechies du mot.

Si nous appliquons cette même procédure au mot *amante* ("amant", mais aussi "amante", comme nom et adjectif), nous aurons la matrice suivante:

ms	fs	mp	fp
e	e	i	i

Table 2

qui décrit la morphologie de ces quatre termes:

amante ms

amante fs

amanti mp

amanti fp

pour lesquels les formes masculines et féminines sont en rapport d'homographie, au singulier comme au masculin.

Cette méthode de décomposition morphologique est appliquée à tous les mots flexionnels, y compris les verbes, pour lesquels les formes fléchies à prévoir sont en nombre plus grand que pour les autres catégories grammaticales. Par exemple, pour un verbe comme *amare* (aimer) qui, dans les grammaires classiques de l'italien, est inséré dans la première conjugaison, la matrice binaire de flexion, aura la structure suivante:

	1 ^e sing.	2 ^e sing.	3 ^e sing.	1 ^e plur.	2 ^e plur.	3 ^e plur.
Ind. prés.	<i>3o</i>	<i>3i</i>	<i>3a</i>	<i>3iamo</i>	<i>3ate</i>	<i>3ano</i>
Ind. imp.	<i>3avo</i>	<i>3avi</i>	<i>3ava</i>	<i>3avamo</i>	<i>3avate</i>	<i>3avano</i>
Ind. pass. sim.	<i>3ai</i>	<i>3asti</i>	<i>3ò</i>	<i>3ammo</i>	<i>3aste</i>	<i>3arano</i>
Ind. fut. prés.	<i>3erò</i>	<i>3erai</i>	<i>3erà</i>	<i>3eremo</i>	<i>3erete</i>	<i>3eranno</i>
Impératif	-	<i>3a</i>	<i>3i</i>	<i>3iamo</i>	<i>3ate</i>	<i>3ino</i>
Subj. prés.	<i>3i</i>	<i>3i</i>	<i>3i</i>	<i>3iamo</i>	<i>3iate</i>	<i>3ino</i>
Subj. imp.	<i>3assi</i>	<i>3assi</i>	<i>3asse</i>	<i>3assimo</i>	<i>3aste</i>	<i>3assero</i>
Cond.	<i>3erei</i>	<i>3eresti</i>	<i>3erebbe</i>	<i>3eremmo</i>	<i>3ereste</i>	<i>3erebbero</i>
	ms	fs	mp	fp		
P. prés.	<i>3ante</i>	<i>3ante</i>	<i>3anti</i>	<i>3anti</i>		
P. pass.	<i>3ato</i>	<i>3ata</i>	<i>3ati</i>	<i>3ate</i>		
Gérondif	<i>3ando</i>					

Table 3

Grâce à cette dernière table²⁹, nous pouvons voir que la procédure de structuration d'un paradigme flexionnel verbal est formellement différente de celui que nous avons adopté pour les noms comme *casa* ou aussi pour les autres mots flexionnels non verbaux. En fait, à côté de chaque désinence apparaît le nombre de lettres qu'il est nécessaire d'éliminer de l'infinitif du verbe avant d'ajouter la désinence flexionnelle.

²⁹Nous notons que dans le paradigme flexionnel des verbes les formes composées ne sont pas incluses, cela parce que des séquences telles que *avessi amato* (avais aimé) ne peuvent pas faire partie d'un dictionnaire électronique de mots simples, non interrompus par des blancs ou des séparateurs.

Cette notation est insérée pour les routines automatiques de flexion. Pendant la phase de lecture des paradigmes, ces routines utiliseront les chiffres pour localiser les points où, à chaque fois, ils devront couper le mot pour ajouter les désinences. Pour toutes les formes de *am-are*, le nombre de lettre à enlever est au nombre de trois; néanmoins, si nous appliquons cette même méthode de flexion à *mangi-are* (manger), lui aussi de la première conjugaison, nous notons que le paradigme change sensiblement:

	1 ^e sing.	2 ^e sing.	3 ^e sing.	1 ^e plur.	2 ^e plur.	3 ^e plur.
Ind. prés.	3o	3	3a	3amo	3ate	3ano
Ind. imp.	3avo	3avi	3ava	3avamo	3avate	3avano
Ind. pass. sim.	3ai	3asti	3ò	3ammo	3aste	3arono
Ind. fut. prés.	4erò	4erai	4erà	4eremo	4erete	4eranno
Impératif	-	3a	3	3amo	3ate	3no
Subj. prés.	3	3	3	3amo	3ate	3no
Subj. imp.	3assi	3assi	3asse	3assimo	3aste	3assero
Cond.	4erei	4eresti	4erebbe	4eremmo	4ereste	4erebbero
	ms	fs	mp	fp		
P. prés.	3ante	3ante	3anti	3anti		
P. pass.	3ato	3ata	3ati	3ate		
Gérondif	3ando					

Table 4

A partir de la table 4, nous pouvons noter que le nombre de lettres à éliminer de l'infinitif de *mangiare* n'est pas constant comme il l'est pour *amare*. En outre, la deuxième personne du singulier de l'indicatif présent, la troisième personne du singulier de l'impératif et les trois personnes du singulier du subjonctif présent de *mangiare* sont obtenues grâce à la seule élimination de trois lettres, i.e. sans ajouter des désinences.

Un exemple ultérieur de particularités flexionnelles à l'intérieur de la première conjugaison est représenté par le verbe *andare* (aller), habituellement classé comme irrégulier et dont la flexion inclut beaucoup d'allomorphes³⁰:

³⁰ Dans la table flexionnelle du verbe *andare*, la double notation pour certaines voix indique la présence de formes surabondantes, phénomène qui se vérifie lorsque une voix verbale a plus d'une forme fléchie.

	1 ^a sing.	2 ^a sing.	3 ^a sing.	1 ^a plur.	2 ^a plur.	3 ^a plur.
Ind. prés.	<i>6vado, 6vo</i>	<i>6vai</i>	<i>6va</i>	<i>3iamo</i>	<i>3ate</i>	<i>6vanno</i>
Ind. imp.	<i>3avo</i>	<i>3avi</i>	<i>3ava</i>	<i>3avamo</i>	<i>3avate</i>	<i>3avano</i>
Ind. pass. sim.	<i>3ai</i>	<i>3asti</i>	<i>3ò</i>	<i>3ammo</i>	<i>3aste</i>	<i>3arono</i>
Ind. fut. prés.	<i>3rò</i>	<i>3rai</i>	<i>3rà</i>	<i>3remo</i>	<i>3rete</i>	<i>3ranno</i>
Impératif	-	<i>6va, 6va', 6vai</i>	<i>6vada</i>	<i>3iamo</i>	<i>3ate</i>	<i>6vadano</i>
Subj. prés.	<i>6vada</i>	<i>6vada</i>	<i>6vada</i>	<i>3iamo</i>	<i>3iate</i>	<i>6vadano</i>
Subj. imp.	<i>3assi</i>	<i>3assi</i>	<i>3asse</i>	<i>3assimo</i>	<i>3aste</i>	<i>3assero</i>
Cond.	<i>3rei</i>	<i>3resti</i>	<i>3rebbe</i>	<i>3remmo</i>	<i>3reste</i>	<i>3rebbero</i>
	ms	fs	mp	fp		
P. prés.	<i>3ante</i>	<i>3ante</i>	<i>3anti</i>	<i>3anti</i>		
P. pass.	<i>3ato</i>	<i>3ata</i>	<i>3ati</i>	<i>3ate</i>		
Gérondif	<i>3ando</i>					

Table 5

Ce dernier paradigme est très différent soit de celui de *amare* soit de celui de *mangiare*. Cela nous amène à affirmer que si pour trois verbes apparemment d'une même conjugaison nous avons trouvé des paradigmes flexionnels différents, il sera nécessaire d'étudier le comportement de chaque verbe pour définir les autres modèles possibles. Cette méthode permettra d'éviter erreurs et fourvoiements, dans des cas où une première analyse semblerait parvenir à une classification aisée.

Les paradigmes obtenus seront progressivement dénombrés et insérés avec les étiquettes grammaticales à côté des entrées dont ils décriront la flexion. Les codes alphanumériques qui en dériveront, ainsi appelés parce formés de lettres et de nombres, auront une relation univoque non ambiguë avec les paradigmes de flexion corrélés, i.e. à un code alphanumérique correspondra un seul paradigme. Néanmoins, en continuant dans l'étude morphologique des lemmes, nous noterons que souvent un seul modèle flexionnel, i.e. un seul code alphanumérique, pourra décrire le comportement de plusieurs unités lexicales et pourra être assigné à des ensembles de mots ayant les mêmes caractéristiques morphologiques. Par exemple, le modèle relatif à *casa* sera aussi valide pour les noms *palestra* (palestre), *finestra* (fenêtre), *pala* (pelle), ou encore pour des adjectifs comme *piovana* (pluviale), *ananassa* (ayant la forme d'un ananas). Le modèle d'*amante* sera valide pour *lestofante* (filou) comme nom et adjectif, et pour l'adjectif *ammaliante* (enchantant). En même temps, le modèle de *mangiare* sera aussi valide pour *bruciare* (brûler). En conclusion, les mots d'une langue seront subdivisés en ensembles différents sur la base de leurs caractéristiques flexionnelles. Donc, les mots suivants:

amante

amare

andare

ammaliante

casa

finestra

lestofante

mangiare

pala

palestra

piovana

dans un dictionnaire électronique seront étiquetés comme suit:

*amante,A79*³¹
amante,N79
amare,V3
andare,V5
ammaliante,A79
casa,N41
finestra,N41
lestofante,A79
lestofante,N79
mangiare,V4
pala,N41
palestra,N41
piovana,A41

En même temps, des mots tels que:

apparentemente (apparentement)
su (sur)

seront étiquetés comme suit:

apparentemente,AVV
su,PREP

³¹ Il est possible de noter que les lemmes et leurs codes respectifs sont séparés par des virgules, qui ont la fonction de séparateurs de champ. Dans une base de données telle qu'un dictionnaire électronique, qui sera traité automatiquement par les routines de flexion, l'utilisation des séparateurs de champ doit être homogène, i.e. chaque séparateur doit marquer toujours la limite entre les même champs, dans ce cas la limite entre le lemme et le code alphanumérique.

Nous pouvons noter qu'un code formé seulement par des lettres, c'est-à-dire seulement alphabétique, sert à indiquer que l'entrée n'est pas flexionnelle, ou mieux qu'elle est invariable.

Un type d'étiquettes ultérieur qui peuvent être assignées aux entrées d'un dictionnaire mais qui ne sont pas morphologiques, concerne, pour les verbes, les caractéristiques grammaticales comme la transitivité, l'intransitivité et les auxiliaires utilisables. Au contraire, sur la base des qualités spécifiques de leurs référents, les substantifs peuvent être sémantiquement subdivisés et étiquetés comme humains, non-humains et animés. Pour un verbe comme *andare*, on utilisera l'étiquette *i* pour indiquer qu'il est intransitif; les noms comme *poliziotto* (policier) seront étiquetés avec *um* (abréviation du mot *umano*, humain), *cavallo* (cheval) sera *anim* (abréviation du mot *animato*, animé), tandis que *aria* (air), *casa* et *sedia* (chaise) seront étiquetés comme *non um* (non-humains). Ces étiquettes seront extrêmement importantes pour l'analyse textuelle automatique et le parsing du logiciel Intex®.

Les entrées de la liste de lemmes ainsi structurées et équipées des codes alphabétiques et alphanumériques forment ce que nous appelons le lexique étiqueté d'une langue, et aussi le DELAS ou dictionnaire électronique des mots simples. Par application de la flexion automatique, du DELAS nous créons le DELAF ou dictionnaire électronique des mots fléchis, dont les entrées, comme nous le verrons, sont équipées d'étiquettes relatives aux caractéristiques des mode, temps, personne et nombre pour les verbes, et des genre et nombre pour les entrées flexionnelles non verbales.

Nous avons décrit la genèse d'un dictionnaire électronique mais, avant de passer à d'autres questions, il faudra revenir sur le contenu des tables précédentes, de (3) à (5), qui mettent en évidence quelques particularités morphologiques des verbes italiens. A l'intérieur de la même première conjugaison classique en *-are*, nous avons pu individualiser trois modèles différents pour *amare*, *mangiare* et *andare*, i.e. des verbes que souvent les grammaires ou les manuels du commerce insèrent dans un seul modèle de conjugaison, en justifiant les différentes réalisations sur la base d'aspects phonétiques

spécifiques³². Probablement, les ouvrages sur papier ne sont pas concernés par la possibilité de créer dans le domaine de la morphologie des modèles heuristiques taxinomiquement descriptifs. Néanmoins, ces ouvrages donnent des descriptions superficielles de phénomènes qui sont plutôt complexes. A ce propos, il faut faire quelques considérations:

- la méthode consolidée de décomposition morphologique, adoptée par les grammaires classiques pour établir à quelle conjugaison appartient un verbe donné, prévoit la subdivision en racine, voyelle thématique et désinence, i.e. *amare* est subdivisé en *am-a-re*, *mangiare* en *mangi-a-re*, *andare* en *and-a-re* et ainsi de suite. Comme instruction supplémentaire, on indique que la voyelle thématique "*peut être présente aussi dans d'autres formes*"³³ outre que dans celle de l'infinitif. De toute façon, cette méthode n'indique pas les formes et leur nombre, laissant le lecteur dans le doute et n'indiquant pas non plus, pour des formes comme *mangiamo* (mangeons) et *amiamo* (amions), où est la voyelle thématique et où commence la désinence;
- même si notre modèle flexionnel de *amare* (table 3) correspond à celui de la première conjugaison présentée dans les grammaires classiques et dans les manuels, il n'est pas possible de dériver de ceci le modèle d'autres verbes, toujours de la première conjugaison, comme par exemple *mangiare*, à moins de le modifier sensiblement. Concrètement, cela peut signifier deux choses: *amare* et *mangiare* ne font peut-être pas partie de la même conjugaison, parce qu'ils ont des caractéristiques morphologiques différentes; ou alors, il est nécessaire d'établir des modèles flexionnels supplémentaires pour décrire ces verbes de la première conjugaison qui sont effectivement différents d'*amare*, i.e. il est nécessaire d'éviter de diluer les différences et de fournir des modèles descriptifs incomplets. Une description précise de la flexion des verbes italiens ne peut pas être effectuée en subdivisant les voix en une partie qui ne change jamais et en deux parties qui

³² Voir aussi Serianni, L. 1988, XI-51, XI-52, XI-53, XI-70 et XI-71.

³³ Voir Serianni L. ouvrage cité, XI-51 (c'est nous qui traduisons).

changent, c'est-à-dire la voyelle thématique et la désinence, et un modèle descriptif valable doit savoir prévoir les mutations précises soit dans une partie, soit dans les autres;

- la méthode que nous avons illustrée a le but de décrire exhaustivement le système flexionnel de l'italien, tandis que pour les grammaires classiques ce but peut justement ne pas être crucial. Notre approche doit tenir compte de chaque particularité, pour arriver à individualiser une quantité probablement très élevée de paradigmes. En tout cas, il faut ne pas oublier que les "exceptions" flexionnelles de la langue italienne ne se produisent pas à cause de la méthode descriptive que nous adoptons ici; plutôt que grâce à celle-ci elles sont mieux mises en évidence. D'ailleurs, les données produites par son application ne peuvent pas être ignorées. Pour cette raison, pour la première conjugaison de l'italien, en utilisant comme modèle le paradigme de *amare*, nous avons dû vérifier l'existence de 2683 verbes "anormaux", sur un nombre total de 8507 formes en *-are* et sur les 10659 lemmes verbaux de l'italien. Pour rendre compte de ces différences, nous avons subdivisé la conjugaison de *-are* en dix-sept paradigmes flexionnels distincts, un pour la flexion de *amare* et seize pour les "exceptions" et les "irréguliers". Dans cette dernière classe rentrent des verbes hautement utilisés, comme *navigare* (naviguer), *nevicare* (neiger), *spaccare* (fendre) ou *tagliare* (couper). De notre point de vue, un tel nombre de "anomalies" met à zéro le concept d'exception, en lui donnant la même valeur que la "norme", donc en demandant une description également soignée et approfondie.

Notre méthode descriptive a donc le but de se délivrer de la tendance au *do it yourself* des grammaires classiques et d'être un "miroir fidèle" des différences flexionnelles de l'italien, transposant notre attention sur des aspects aplatis par des modèles méthodologiquement imparfaits.

2.2 LE DICTIONNAIRE FLECHI

Dans le sous-chapitre précédent nous avons décrit les phases essentielles de la construction du DELAS, tandis que dans celui-ci nous verrons comment les informations grammaticales associées aux lemmes de ce dictionnaire permettent de produire les formes fléchies correspondantes. Cette opération est accomplie en transformant les modèles de flexion en transducteurs à états finis, qui associés à des routines informatiques, comme on l'a déjà vu, fléchissent le DELAS en DELAF³⁴.

Grâce à sa systématité formelle, un modèle de flexion comme celui décrit pour *casa, N41* peut être converti en expression régulière et donc devenir un transducteur à états finis. La conversion des expressions régulières se fait à l'aide d'une instruction du logiciel Intex® qui, à partir de l'image graphique d'un modèle tel que celui de la figure suivante:

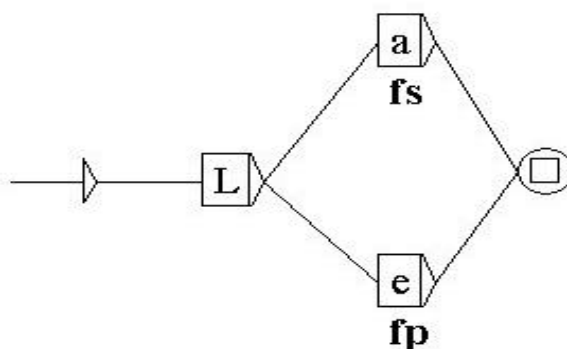


Figure 1

élabore automatiquement la routine à appliquer aux lemmes qui ont le paradigme de cette classe flexionnelle. Dans la figure 1, l'instruction "L" insérée dans l'état initial du graphe indique qu'il est nécessaire d'effacer le dernier caractère du mot simple à fléchir.

³⁴ Voir Silberztein, M. 1997.

Un nombre plus haut d'instructions "L" dans l'état initial correspond à un nombre équivalent de caractères à effacer: l'instruction "LL" efface donc deux caractères, "LLL" en efface trois, et ainsi de suite. Avec ces effacements nous obtenons des racines, comme par exemple *cas-*, *pal-*, *sedi-*, auxquelles pour effectuer la flexion le transducteur associera les désinences que nous trouvons dans les autres états du graphe. Le graphe de la figure 1 associera donc aux racines la désinence *-a* pour le féminin singulier et la désinence *-e* pour le féminin pluriel.

Pour fléchir le DELAS en DELAF, il est nécessaire de créer un transducteur pour chaque classe flexionnelle individualisée. Les transducteurs les plus complexes à réaliser sont ceux pour les verbes, à cause du haut nombre de voix fléchies déjà mises en évidence dans les tables de (3) à (5).

Une fois terminée la création des transducteurs, une option d'INTEX® permet de les appliquer au DELAS tous simultanément. A partir des formes simples indiquées dans le paragraphe précédent, nous obtiendrons automatiquement les formes fléchies suivantes:

ama, amare. V: Q2s
ama, amare. V: X3s
amai, amare. V: J1s
amammo, amare. V: J1p
amando, amare. V: G
amano, amare. V: X3p
amante, amante. A: fs
amante, amante. A: ms
amante, amante. N: fs
amante, amante. N: ms
amante, amare. V: Zms: Zfs
amanti, amante. A: fp
amanti, amante. A: mp
amanti, amante. N: fp
amanti, amante. N: mp
amanti, amare. V: Zmp: Zfp
amar, amare. V: L
amare, amare. V: I
amarono, amare. V: J3p
amasse, amare. V: H3s
amassero, amare. V: H3p
amassi, amare. V: H1s
amassi, amare. V: H2s
amassimo, amare. V: H1p
amaste, amare. V: H2p
amaste, amare. V: J2p
amasti, amare. V: J2s
amata, amare. V: Ufs
amate, amare. V: Q2p
amate, amare. V: Ufp
amate, amare. V: X2p
amati, amare. V: Ump
amato, amare. V: Ums
amava, amare. V: Y3s
amavamo, amare. V: Y1p
amavano, amare. V: Y3p
amavate, amare. V: Y2p
amavi, amare. V: Y2s
amavo, amare. V: Y1s

amerai, amare. V: K2s
ameranno, amare. V: K3p
amerebbe, amare. V: F3s
amerebbero, amare. V: F3p
amerei, amare. V: F1s
ameremmo, amare. V: F1p
ameremo, amare. V: K1p
amereste, amare. V: F2p
ameresti, amare. V: F2s
amerete, amare. V: K2p
amerà, amare. V: K3s
amerò, amare. V: K1s
ami, amare. V: Q3s
ami, amare. V: W1s
ami, amare. V: W2s
ami, amare. V: W3s
ami, amare. V: X2s
amiamo, amare. V: Q1p
amiamo, amare. V: W1p
amiamo, amare. V: X1p
amiate, amare. V: W2p
amino, amare. V: Q3p
amino, amare. V: W3p
ammaliante, ammaliante. A: fs
ammaliante, ammaliante. A: ms
ammalianti, ammaliante. A: fp
ammalianti, ammaliante. A: mp
amo, amare. V: X1s
amò, amare. V: J3s
andai, andare. V: J1s
andammo, andare. V: J1p
andando, andare. V: G
andante, andare. V: Zms: Zfs
andanti, andare. V: Zmp: Zfp
andar, andare. V: L
andare, andare. V: I
andarono, andare. V: J3p
andasse, andare. V: H3s
andassero, andare. V: H3p

andassi, andare. V: H1s
andassi, andare. V: H2s
andassimo, andare. V: H1p
andaste, andare. V: H2p
andaste, andare. V: J2p
andasti, andare. V: J2s
andata, andare. V: Ufs
andate, andare. V: Q2p
andate, andare. V: Ufp
andate, andare. V: X2p
andati, andare. V: Ump
andato, andare. V: Ums
andava, andare. V: Y3s
andavamo, andare. V: Y1p
andavano, andare. V: Y3p
andavate, andare. V: Y2p
andavi, andare. V: Y2s
andavo, andare. V: Y1s
andiamo, andare. V: Q1p
andiamo, andare. V: W1p
andiamo, andare. V: X1p
andiate, andare. V: W2p
andrai, andare. V: K2s
andranno, andare. V: K3p
andrebbe, andare. V: F3s
andrebbero, andare. V: F3p
andrei, andare. V: F1s
andremmo, andare. V: F1p
andremo, andare. V: K1p
andreste, andare. V: F2p
andresti, andare. V: F2s
andrete, andare. V: K2p
andrà, andare. V: K3s
andrò, andare. V: K1s
andò, andare. V: J3s
casa, casa. N: fs
case, casa. N: fp
finestra, finestra. N: fs
finestre, finestra. N: fp

lestofante, lestofante. A: fs
lestofante, lestofante. A: ms
lestofante, lestofante. N: fs
lestofante, lestofante. N: ms
lestofanti, lestofante. A: fp
lestofanti, lestofante. A: mp
lestofanti, lestofante. N: fp
lestofanti, lestofante. N: mp
mangerai, mangiare. V: K2s
mangeranno, mangiare. V: K3p
mangerebbe, mangiare. V: F3s
mangerebbero, mangiare. V: F3p
mangerei, mangiare. V: F1s
mangeremmo, mangiare. V: F1p
mangeremo, mangiare. V: K1p
mangereste, mangiare. V: F2p
mangeresti, mangiare. V: F2s
mangerete, mangiare. V: K2p
mangerà, mangiare. V: K3s
mangerò, mangiare. V: K1s
mangi, mangiare. V: Q3s
mangi, mangiare. V: W1s
mangi, mangiare. V: W2s
mangi, mangiare. V: W3s
mangi, mangiare. V: X2s
mangia, mangiare. V: Q2s
mangia, mangiare. V: X3s
mangiai, mangiare. V: J1s
mangiammo, mangiare. V: J1p
mangiamo, mangiare. V: Q1p
mangiamo, mangiare. V: W1p
mangiamo, mangiare. V: X1p
mangiando, mangiare. V: G
mangiano, mangiare. V: X3p
mangiante, mangiare. V: Zms: Zfs
mangianti, mangiare. V: Zmp: Zfp
mangiar, mangiare. V: L
mangiare, mangiare. V: I
mangiarono, mangiare. V: J3p

mangiasse, mangiare. V: H3s
mangiassero, mangiare. V: H3p
mangiassi, mangiare. V: H1s
mangiassi, mangiare. V: H2s
mangiassimo, mangiare. V: H1p
mangiaste, mangiare. V: H2p
mangiaste, mangiare. V: J2p
mangiasti, mangiare. V: J2s
mangiata, mangiare. V: Ufs
mangiate, mangiare. V: Q2p
mangiate, mangiare. V: Ufp
mangiate, mangiare. V: W2p
mangiate, mangiare. V: X2p
mangiati, mangiare. V: Ump
mangiato, mangiare. V: Ums
mangiava, mangiare. V: Y3s
mangiavamo, mangiare. V: Y1p
mangiavano, mangiare. V: Y3p
mangiavate, mangiare. V: Y2p
mangiavi, mangiare. V: Y2s
mangiavo, mangiare. V: Y1s
mangino, mangiare. V: Q3p
mangino, mangiare. V: W3p
mangio, mangiare. V: X1s
mangiò, mangiare. V: J3s
palestra, palestra. N: fs
palestre, palestra. N: fp
piovana, piovana. A: fs
piovane, piovana. A: fp
va', andare. V: Q2s
va, andare. V: Q2s
va, andare. V: X3s
vada, andare. V: Q3s
vada, andare. V: W1s
vada, andare. V: W2s
vada, andare. V: W3s
vadano, andare. V: Q3p
vadano, andare. V: W3p
vado, andare. V: X1s

vai, andare. V: Q2s
vai, andare. V: X2s
vanno, andare. V: X3p
vo, andare. V: X1s

Cette liste nous aide à observer que les chaînes du DELAF sont structurées de façon différente de celle du DELAS. En lisant de gauche à droite nous trouvons les éléments suivants:

- le mot dans sa forme fléchie;
- un premier séparateur de champ, c'est-à-dire la virgule ",";
- le mot dans sa forme canonique, c'est-à-dire non fléchie;
- un second séparateur, c'est-à-dire le point ".";
- l'étiquette grammaticale;
- un troisième séparateur, c'est-à-dire le point-virgule ";";
- les informations grammaticales corrélées à la forme fléchie.

Quant aux étiquettes verbales, qui dans le DELAF sont différentes de celles non verbales, elles incluent des sigles relatifs aux modes et aux temps des formes fléchies, que nous listons en ordre alphabétique³⁵:

F = conditionnel présent

G = gérondif présent

H = subjonctif imparfait

I = infinitif présent

J = indicatif passé simple

³⁵ Étant donné que le DELAF est un dictionnaire de mots simples, les étiquettes pour les temps verbaux composés ne sont pas prévues.

K = indicatif futur présent

L = infinitif élide

Q = impératif présent

U = participe passé

W = subjonctif présent

X = indicatif présent

Y = indicatif imparfait

Z = participe présent

En outre, elles comprennent aussi d'autres étiquettes alphanumériques, qui indiquent le nombre des mêmes flexions. L'étiquette:

V:Q2p

indiquera que cette entrée est la deuxième personne plurielle (2p) de l'impératif présent (Q). Nous observons que les étiquettes des entrées du DELAF sont seulement alphabétiques et non alphanumériques comme celles du DELAS. Pendant la phase de flexion automatique, les étiquettes numériques sont effacées parce que leur présence n'est plus nécessaire.

A partir de 15 formes canoniques nous avons obtenu 188 formes fléchies, donc avec un ratio moyen de 1:12,61. Globalement, le ratio entre les versions actuelles du DELAS et du DELAF, pris dans leur totalité, est à peu près de 1:6,61. Ce rapport est calculé sur la base des presque cent trente mille mots canoniques présents dans le DELAS. Cela nous aide à donner une estimation approximative des dimensions du lexique italien, qui inclut actuellement plus de 859.000 mots.

2.3 LE DICTIONNAIRE DE MOTS COMPOSES

Les différences formelles, morphologiques et sémantiques qui existent entre mots simples et mots composés justifient, en terme de lemmatisation, la séparation entre les deux types d'entrées et donc la création de deux dictionnaires différents, où mieux de deux bases de données distinctes. A côté du dictionnaire électronique des mots simples, nous en aurons donc un autre spécifique pour les mots composés. L'élaboration différenciée des deux dictionnaires est essentielle parce qu'elle permettra de diriger l'analyse textuelle automatique vers des structures sémantiques et morphologiques prédéfinies. La réalisation de dictionnaires électroniques spécifiques sert aussi à effectuer des analyses textuelles non ambiguës, i.e. à examiner une seule unité lexicale particulière comme peut l'être *a braccia conserte* (les bras croisés) outre ses éléments simples. Cette séparation sera utile au logiciel d'analyse textuelle automatique qui ne peut utiliser comme moteur linguistique que des bases de données lexicales formellement homogènes.

Un exemple de dictionnaire électronique de mots composés (dorénavant indiqué comme *DELAC*) est celui de la liste suivante:

a capo automatico,PNA+N:ms-+;TRATT TESTI

a commerciale,NA+N:ms-+;NOTAZ

abaco biquinario,NA+N:ms-+;MAT

abasia coreica,NA+N:fs-+;MED

Comme nous pouvons le noter, les chaînes du *DELAC* sont structurées différemment de celles du *DELAS-DELA*F, et elles incluent:

- le mot composé canonique, i.e. non fléchi;
- le séparateur de champ " , " ;

- une série d'étiquettes grammaticales qui indiquent la structure interne du mot composé (par exemple, la série NPN indique que le mot composé est formé par un Nom, suivi d'une Préposition suivie d'un autre Nom);
- le séparateur de champ "+";
- une étiquette grammaticale (N, V et ainsi de suite) qui indique la fonction du mot composé;
- le séparateur de champ ":";
- les informations morphologiques relatives au mot composé (comme déjà indiqué, **m** pour masculin, **f** pour féminin, **s** pour singulier, **p** pour pluriel) suivies par les indications relatives aux autres formes possibles. De cette façon, une étiquette du type **fs-+** indiquera que le mot composé est féminin singulier, qu'il n'a pas de forme masculine correspondante (comme l'indique le signe -) et qu'il peut au contraire avoir une forme féminine plurielle (comme l'indique le signe +);
- le séparateur de champ ";"
- les informations relatives aux domaines lexicaux et sémantiques spécifiques dans lesquels le mot composé est employé (ECON pour économie, FIS pour physique, INF pour informatique, et ainsi de suite). Grâce à la liste précédente, nous pouvons observer que quelques mots composés sont utilisés dans plus d'un domaine.

Pendant la lemmatisation, on étudie toutes les formes fléchies possibles des mots composés, et à l'intérieur de celles-ci les unités lexicales simples qui fléchissent sont étiquetées avec les codes alphanumériques déjà repérés pour les mots simples. Après cette procédure, nous obtenons un dictionnaire dont les entrées ont la structure suivante:

a capo automatico(A87),PNA+N:ms-+;TRATT TESTI

a(N601) commerciale(A79),NA+N:ms-+;NOTAZ

abaco(N10) biquinario(A95),NA+N:ms-+;MAT

abasia(N41) coreica(A87),N,NA+N:fs-+;MED

Avec un contrôle sur les codes et sur les étiquettes morphologiques des mots composés, la routine de flexion automatique localise les formes fléchies à réaliser en appliquant les mêmes procédures vues pour le DELAS, et à partir du DELAC crée le *DELACF*, i.e. le dictionnaire électronique des mots composés fléchis, dont les entrées ont la structure suivante:

a capo automatico,PNA+N:ms-+;TRATT TESTI
a commerciale,NA+N:ms-+;NOTAZ
a commerciali,a commerciale.NA+N:fp-+;NOTAZ
a commerciali,a commerciale.NA+N:fp-+;NOTAZ
a commerciali,a commerciale.NA+N:mp-+;NOTAZ
a maiuscola,a maiuscolo.NA+N:fs++;LING
a maiuscole,a maiuscolo.NA+N:fp++;LING
a maiuscoli,a maiuscolo.NA+N:mp++;LING
a maiuscolo,.NA+N:ms++;LING
a minuscola,a minuscolo.NA+N:fs++;LING
a minuscole,a minuscolo.NA+N:fp++;LING
a minuscoli,a minuscolo.NA+N:mp++;LING
a minuscolo,.NA+N:ms++;LING
abachi biquinari,abaco biquinario.NA+N:mp-+;MAT
abaco biquinario,.NA+N:ms-+;MAT
abasia coreica,.NA+N:fs-+;MED
abasia coreiche,abasia coreica.NA+N:fp-+;MED

Par rapport à celles du DELAC, les chaînes du DELACF ajoutent un séparateur de champ, i.e. le point "." qui sert à distinguer la forme fléchie de la forme canonique. Les chaînes marquées par la séquence de séparateurs ".,." indiquent formellement que l'entrée est au singulier.

A partir d'un mot composé singulier, il est possible de faire dériver en moyenne 4,25 formes fléchies. La taille actuelle du DELAC est de presque vingt-cinq mille entrées, tandis que celle du DELACF est de presque cent mille entrées.

2.4 MOTS, SIGNIFIES ET USAGES DANS LES DICTIONNAIRES ELECTRONIQUES

Par rapport aux dictionnaires électroniques, dans les pages suivantes nous affronterons le sujet de la crédibilité des sources lexicales de repérage, des modalités de mise à jour, des critères à adopter pour l'insertion des nouvelles entrées et aussi de l'évaluation des néologismes potentiels. Nous verrons aussi comment les codes de flexion du DELAS peuvent être partiellement porteurs des caractéristiques syntaxiques et sémantiques des entrées et aussi comment ils sont aptes à résumer quelques aspects des gloses des dictionnaires papier.

Les caractéristiques applicatives du DELAS et du DELAF rendent nécessaires l'insertion dans la liste des lemmes du nombre le plus haut possible de voix, et nous verrons que cette nécessité sera ultérieurement mise en évidence par les procédures d'analyse textuelle automatique. Néanmoins, il faudra faire attention à ne pas inclure dans le dictionnaire n'importe quel lemme dont on a vérifié l'usage, sans appliquer aucun critère sélectif et en présumant ainsi de sauvegarder les finalités et les caractéristiques descriptives déjà exposées à propos des dictionnaires électroniques. L'adjonction indiscriminée des lemmes peut être une erreur capable d'invalider la fonctionnalité des dictionnaires électroniques. Pour éviter cela, il est nécessaire d'effectuer une analyse attentive des sources de repérage des lemmes, surtout pour celles qui concernent les néologismes.

2.4.1 Comment choisir les lemmes

La liste de lemmes du dictionnaire papier reste l'une des sources les plus importantes et crédibles pour la réalisation des dictionnaires électroniques; mais nous avons déjà dit qu'elle ne peut pas être la seule source, parce qu'elle n'inclut pas tous les lemmes utilisés dans une langue. La non-exhaustivité du dictionnaire papier impose donc de chercher ailleurs les lemmes qu'elle n'inclut pas. Cet aspect fait indirectement naître le problème de la crédibilité des sources nouvelles de repérage.

En termes médiatiques, il est possible de subdiviser ces sources en deux grands groupes, sur la base des instruments de divulgation et communication qu'ils utilisent. Le premier groupe est celui de la langue écrite, et il inclut tous les médias qui utilisent des textes imprimés³⁶; le deuxième groupe est celui de la langue parlée et comprend tous les médias ou les situations communicatives dans lesquelles on utilise des messages oraux. Certes, les modalités de repérage des nouveaux mots seront différentes pour les deux groupes, vu qu'avec les textes imprimés la réception et le contrôle du contenu linguistique seront plus simples, tandis que pour les messages oraux elles devront être mémorisées (d'une façon ou d'une autre), avant d'être contrôlées. Pour cette raison, pendant une première phase de repérage, il sera convenable de favoriser la recherche sur des textes écrits, qui se présentera en fait comme plus simple.

Une source écrite de repérage lexical est hautement crédible quand:

- elle est réalisée et/ou révisée par des experts du secteur dont elle traite;
- la révision à laquelle elle sera soumise est constante dans le temps, en ce qui concerne la forme comme le contenu;

³⁶ En particulier, pour les mots composés, les dictionnaires de spécialité sur papier, i.e. ces dictionnaires qui s'occupent de secteurs lexicaux et sémantiques spécifiques, portent une grande importance à ce sujet parce qu'ils incluent des tecnicismes et des néologismes qu'un dictionnaire ordinaire sur papier n'a pas raison de lister.

- elle est destinée à un public non restreint, c'est-à-dire qu'elle a un niveau de vulgarisation suffisant;
- elle est dotée de continuité temporelle, i.e. elle n'est pas hors du temps et a la possibilité de se représenter, culturellement et médiatiquement, aux locuteurs d'une langue donnée.

Ces quatre conditions générales garantissent la scientificité de la source et de la langue adoptée, la compréhensibilité et la vaste diffusion des mots utilisés et aussi l'apport lexical non hors du temps à la langue d'usage. Elles résolvent donc la plupart des problèmes liés à la crédibilité des sources. Implicitement, elles indiquent aussi quel doit être le rôle du linguiste pendant la phase d'analyse des sources, et en délimitent les tâches, soit comme locuteur de langue maternelle, capable de donner des jugements sur les qualités linguistiques générales des mots utilisés, soit comme expert, donc comme quelqu'un qui, par rapport aux nouveaux mots repérés, doit valider des facteurs tels que signifié, compréhensibilité, dicibilité et conformité aux critères sémiologiques de la langue d'usage.

Les nouveaux mots avec lesquels les dictionnaires électroniques pourront être mis à jours sont théoriquement subdivisibles comme suit:

- les termes techniques non néologiques;
- les termes techniques néologiques;
- les termes non techniques non néologiques;
- les termes non techniques néologiques.

Plus spécifiquement, les termes non néologiques seront représentés par les mots déjà existant dans une langue qui n'ont pas été insérés dans les dictionnaires

électroniques. Au contraire, les néologismes seront tous ces mots créés sur la base des exigences culturelles spécifiques d'une langue. Pour les termes techniques, (i.e. ces mots souvent composé qui sont utilisés pour indiquer les entités appartenant à des domaines spécifiques du savoir), l'analyse devra être focalisée sur des aspects spécifiquement terminologiques. Pour les termes non techniques, dans leur forme soit simple soit composée, il faudra évaluer l'impact morphologique, grammatical et sémantique sur la langue d'usage.

Les néologismes non techniques sont les résultats de la créativité d'une langue, donnent forme aux nouveautés sociales et coutumières et ont une fréquence d'usage élevée pendant leur existence, qu'elle soit longue ou brève. Le mot *Sessantottino*³⁷ (Soixante-huitard) est un bon exemple de néologisme non technique qui a longtemps vécu, tandis que *paninaro*³⁸ représente un cas exemplaire de disparition précoce. Ces mots sont presque toujours lemmatisés dans les dictionnaires papier, qui néanmoins s'emploient à les effacer de la liste des lemmes une fois qu'ils deviennent obsolètes. Au contraire, les dictionnaires électroniques tendent à les garder tous, surtout pour sauvegarder la perspective diachronique³⁹ qui est typique de l'analyse textuelle.

Avant d'intégrer dans un dictionnaire électronique un néologisme non technique, il est nécessaire d'effectuer une analyse des processus sémiologiques de sa création, qui doivent être analysés et adoptés comme critères discriminatoires. A ce propos, nous dirons qu'un néologisme non technique pourra être inséré dans un dictionnaire électronique lorsque:

- il est d'usage commun;
- il est sémantiquement prégnant, i.e. il a un signifiant et un signifié spécifique;

³⁷ Personne qui a été politiquement actif à la fin des années soixante. On l'emploie pour indiquer les gens appartenant à la génération de Mai soixante-huit.

³⁸ Créé à partir de *panino*, (sandwich), pour indiquer une catégorie sociale de jeune gens qui, pendant les années quatre-vingts, fréquentaient les fast-foods et s'habillaient avec des chaussures Timberland® et des blousons rembourrés de plumes d'oie.

³⁹ Par exemple, le mot *cecoslovacco* (tchécoslovaque), qui n'est plus utilisable par rapport à une nationalité spécifique, ne peut pas être effacé d'un dictionnaire électronique parce qu'il pourrait de toute façon paraître dans un texte plus ancien que l'on voudrait analyser automatiquement.

- sa création est faite sur la base des règles phonétiques et de dérivation morphologique spécifiques d'une langue donnée;
- il comble une lacune lexicale⁴⁰, en termes soit de catégorie grammaticale, soit sémantiques;
- il est non ambigu à la lecture, c'est-à-dire qu'il peut être interprété grâce à la simple connaissance encyclopédique du locuteur, et il n'induit pas formellement d'interprétations erronées.

Par leur caractère essentiel, ces critères limitent fortement le choix des néologismes non techniques aptes à mettre à jour le dictionnaire électronique, et ils aident à individualiser surtout les mots qui ne sont pas à insérer. En outre, leur application produit des résultats efficaces, comme le démontrent les exemples suivants de néologismes non techniques qui violent les normes que nous venons d'exposer et qui n'ont pas eu de chance dans la lexicographie classique de l'italien, et, par reflet, aussi dans les dictionnaires électroniques:

- *htmlista* (prononcé [ak:atiem:el:ista]), en référence aux utilisateurs du langage *html*;
- *inputazione*, dérivé de l'anglais *input* et en référence au *data entry*;
- *pulastina*, en référence à un objet utilisé pour nettoyer (*puli-*, de *pulire*) une petite hampe (*asta* en italien);

⁴⁰ Les lacunes lexicales se vérifient lorsqu'il n'est pas possible d'exprimer un concept déjà existant avec un seul mot, ce qui rend nécessaire l'utilisation de paraphrases. Par exemple, en partant du modèle morphologique du mot *media* (média), *mediale* (médial) et *medialità* (médialité), nous définirons comme lacune lexicale l'absence de mots tels que *cosale* (de chose) et *cosalità* (l'essence des choses) dérivable de *cosa* (chose).

Si nous analysons ces mots, nous observons que *htmlista* contient une séquence de phonèmes exotiques pour la langue italienne, i.e. *html* qui est un acronyme d'origine anglaise, prononcé lettre par lettre, auquel on a ajouté le suffixe nominal et adjectival – *ista*. Ce mot a été utilisé pendant une période de temps très brève mais il est rapidement devenu obsolète parce qu'il était difficilement prononçable et lisible. A sa place, on utilise aujourd'hui les mots composés *operatore Web* (opérateur Web) et *programmatore html* (programmeur html).

La non lemmatisation de *inputazione* est due à deux facteurs précis: en premier lieu, la présence de la séquence de lettres *np*, elle aussi exotique pour l'italien qui utilise les phonèmes *mp*⁴¹ qui ont une prononciation identique; en deuxième lieu, son faible apport sémantique, vu qu'il fait référence à une activité pour laquelle on peut aussi utiliser trois unités lexicales – un emprunt anglo-saxon composé, i.e. *data entry*, le correspondant calque italien *immissione dati* et un emprunt anglo-saxon simple, i.e. *input*. Finalement, le mot *puliastina*, qui n'est pas à lemmatiser, est sémantiquement et formellement ambigu parce qu'il ressemble à quelques termes du domaine de la chimie et de la biologie, tels que amino-pyridine ou élastine, et donc il provoque une interprétation erronée. En outre, il viole les règles de dérivation morphologique de l'italien, vu qu'il est un composé monorhématique, formé par la voix verbale *pulisci* (nettoies) et le substantif *astina* (petite hampe) diminutif de *asta*. Les composés monorhématiques sont plutôt fréquents en italien et incluent des mots comme *aspirapolvere* (aspirateur à poussière), *macinacaffè* (moulin à café) et *tagliaerba* (tondeuse à gazon). Les anomalies formelles de *puliastina* par rapport aux autres composés monorhématiques sont principalement de deux sortes: l'apocope immotivée de la dernière syllabe de *pulisci* – qui ne se vérifie pas avec *aspira*, *macina* ou *taglia* – et l'utilisation d'un diminutif qui dans les autres mots produirait des résultats bizarres comme *macinacaffettuccio*, *tagliaerbetta* et *aspirapolverina*.

⁴¹ Fait exception la graphie de quelques noms propres comme *Gianpaolo* (Jean-Paul) ou *Gianpiero* (Jean-Pierre), obtenu par verbalisation unique de l'hypocoristique *Gianni* avec *Paolo* et *Pietro*.

2.4.2 La sémantique dans le dictionnaire électronique

Le lexique-grammaire a démontré que l'ambiguïté des unités lexicales simples se réduit sensiblement à l'intérieur de contextes phrastiques syntaxiquement et sémantiquement corrects. Dans n'importe quelle phrase, si elle est correcte, il est possible d'établir le signifié d'un mot sur la base des autres avec lesquels il co-occure. En ce sens, il est licite d'affirmer que les contextes phrastiques sont justement porteurs d'informations morphologiques spécifiques et syntaxiques qui réduisent sensiblement l'ambiguïté de mots même polysémiques. Par exemple, dans la phrase:

(I) *La squadra ha dichiarato la sua disponibilità a giocare di sabato*

(L'équipe s'est déclarée disponible à jouer le samedi)

l'occurrence du verbe *dichiarare* (déclarer) permet d'interpréter le substantif *squadra* seulement dans le sens de "*insieme di giocatori o atleti*" (ensemble de joueurs ou d'athlètes) et non aussi de "*strumento a forma di triangolo rettangolo, atto a tracciare le perpendicolari e le parallele ad una linea retta data*" (instrument en forme de triangle rectangle apte à tracer les perpendiculaires et les parallèles à une ligne droite donnée)⁴². Donc, individualiser une méthode descriptive adéquate à tous les contextes phrastiques dans lesquels un mot peut figurer, et ensuite formaliser cette méthode, veut dire au moins théoriquement permettre aussi à un ordinateur de donner pendant le parsing d'un texte des interprétations univoquement correctes.

Le succès d'une analyse syntaxique automatique de ce type est fortement dépendant de la création d'un dictionnaire électronique dans lequel les unités lexicales sont étiquetées avec des codes eux aussi interprétables de façon univoque. Néanmoins, au dehors des contextes phrastiques et surtout pendant la phase de classification et lemmatisation, les unités lexicales simples se prêtent à plusieurs interprétation, comme

⁴² En italien, *squadra* signifie aussi *équerre*. Les deux définitions ont été reprises de Dogliotti, M., Rosiello, G., 1996 et traduites par l'auteur.

nous l'avons vu pour le mot *ufficio*. En outre, quand nous avons traité de la récupération des informations sur le Web, nous avons aussi établi que l'ambiguïté lexicale, morphologique et grammaticale d'un mot simple, analysé en dehors d'un contexte d'occurrence, ne peut pas être automatiquement résolue par un ordinateur, au moins pendant la phase de lecture automatique d'un texte. Pour tenter de donner une solution au moins partielle à ces problèmes, nous démontrerons que les codes de flexion associés aux lemmes d'un dictionnaire électronique peuvent être utiles pour effectuer une première subdivision du lexique d'une langue en plusieurs portions sémantiques. Le principe de séparation synchronique des signifiés et des usages d'un mot servira à justifier les différenciations des entrées, également pour des lemmes qui ont de fortes analogies formelles et sémantiques ou des particularités relatives à des interprétations opposées les unes aux autres, comme par exemple celles qui sont littéraires *versus* celles qui sont figurées.

Nous avons vu que pendant les phases de structuration et de mise à jour, l'assignation d'un code de flexion à une entrée d'un dictionnaire électronique est faite sur bases empiriques, c'est-à-dire en analysant les propriétés morphologiques et syntaxiques des mots étudiés à l'intérieur de différents contextes phrastiques. Par exemple, en étudiant les occurrences du mot italien *matricolare* (matriculaire et immatriculer), nous observerons qu'il peut être utilisé comme adjectif ou comme verbe. Pour cette raison, dans le DELAS nous insérerons deux entrées différentes pour ce mot:

matricolare,A79

matricolare,V3

En présence de cas semblables, i.e. avec des lemmatisations multiples, les codes de flexion assument une fonction similaire à celle des gloses des dictionnaires papier, i.e. ils donnent des informations sur les différences grammaticales des lemmes. La lemmatisation de *matricolare* n'est pas particulièrement complexe, vu que ce mot nécessite *seulement* deux entrées. Les difficultés augmentent quand une unité lexicale est polysémique à l'intérieur d'une même catégorie grammaticale. Par exemple, l'analyse

morphologique et syntaxique du substantif *accordatore* (accordeur) servira à mettre en évidence qu'on peut indiquer par ce mot soit l'outil employé pendant l'accordage des différents instruments musicaux, soit la personne qui accorde ces mêmes instruments. Dans le premier cas, *accordatore* sera un lemme non-humain masculin. Dans le deuxième cas, il sera un nom humain, avec une forme féminine du singulier en *accordatrice* (accordeuse). Pour cette raison, dans le DELAS nous aurons les lemmatisations suivantes:

accordatore,N5

accordatore,N81

où *N5* indique le substantif non-humain et *N81* celui qui est humain. En italien, un pareil redoublement est souvent nécessaire avec les substantifs qui se terminent en *-ore*, comme nous pouvons l'observer dans la liste suivante:

accordatore,N5

accordatore,N81

accumulatore,N5 (accumulateur)

accumulatore,N81 (accumulateur)

trasportatore,N5 (transporteur)

trasportatore,N81(rouleur)

vogatore,N5 (machine à ramer)

vogatore,N81(rameur)

Le code *N81* opposé au code *N5* permet de distinguer entre lemmes avec référents humains et lemmes avec référents non-humains.

Un type ultérieur de lemmatisation double peut se vérifier avec des substantifs dont l'usage dépend de caractéristiques référentielles et sémantiques. Par exemple, dans le DELAS le mot *centauro* (centaure) est lemmatisé:

centauro,N110

centauro,N7

Le paradigme flexionnel du code prévoit les formes:

centauro: substantif masculin singulier

centaura o *centauressa*: substantif féminin singulier

centauri: substantif masculin pluriel

centaure o *centauresse*: substantif féminin pluriel

tandis que le paradigme *N7* a seulement les formes suivantes:

centauro: substantif masculin singulier

centauri: substantif masculin pluriel

On en déduit que *centauro,N110* est utilisé pour indiquer la créature mythologique, tandis que avec *centauro,N7* on indique métaphoriquement un motocycliste (mâle ou femelle). En ce cas, les deux codes en opposition nous permettent de distinguer entre signifiés et usages différents.

En utilisant cette méthode de différenciation, il est possible de rendre transparents la plupart des termes polysémiques qui peuvent provoquer des interprétations ambiguës. Néanmoins, il y a des mots dont l'ambiguïté ne peut pas être résolue, comme par exemple:

cubista, N70 (cubiste)

terme utilisé pour indiquer soit les prosélytes du cubisme, soit les personnes qui ont pour profession de danser sur des cubes dans les discothèques. En effet, le paradigme flexionnel N70 prévoit les formes suivantes:

cubista: substantif masculin et féminin singulier

cubisti: substantif masculin pluriel

cubiste: substantif féminin pluriel

mais ne donne aucune indication utile pour différencier les deux usages, tous les deux à référent humain. Dans ce cas, les indications fournies par le dictionnaire électronique peuvent paraître moins précises par rapport à celles présentes dans les dictionnaires papier, mais il faut dire que pour le parsing automatique il n'est pas important d'avoir des informations précises sur les nuances de sens des mots, tandis qu'il l'est en ce qui concerne leurs traits sémantiques essentiels, vu que ce sont ces derniers qui conditionnent et gouvernent les règles de co-occurrence et les restrictions de sélection. Dans le cas de *cubista*, par exemple, les deux interprétations possibles ne créent pas de phrases syntaxiquement incorrectes, comme on peut le vérifier dans la phrase:

(II) *I cubisti danzano al ritmo della musica*

(Les cubistes dansent au rythme de la musique)

En réalité, dans cette phrase, la distinction possible entre les deux signifiés de *cubista* n'aurait pas d'effet sur la régularité morphologique, syntaxique et sémantique du contexte.

CHAPITRE III. DICTIONNAIRE ELECTRONIQUE VERSUS DICTIONNAIRE PAPIER: LE ROI EST NU!

En comparant le dictionnaire électronique avec celui sur papier, jusqu'à présent nous avons mis en évidence des buts applicatifs et des appareillages méthodologiques profondément dissemblables, qui produisent de leur côté des modalités de structuration et de compilation extrêmement différentes. Dans cette partie de conclusion, nous parlerons du dictionnaire papier en tant qu'instrument descriptif du lexique d'une langue, et nous verrons qu'en ce qui concerne le domaine qui lui est commun avec le dictionnaire électronique, c'est-à-dire l'analyse et la classification des mots, il est en retard en ce qui concerne la portée du corpus dépouillé et la crédibilité de l'enquête développée. Nous chercherons à expliquer quelles sont les raisons qui amènent à ces conditions⁴³, en soulignant que, dans les dictionnaires papier, la création et la description correcte d'une liste de lemmes complète et actualisée, semblent des buts à poursuivre plutôt que des objectifs acquis. En outre, nous vérifierons comment les normes lexicographiques spécifiques ne sont pas appliquées constamment, i.e. le dictionnaire papier viole souvent les règles qu'il s'impose de façon autonome, ce qui a pour conséquence parfois d'égarer ses lecteurs.

⁴³ Voir Gross, M. 1989.

3.1 ABSENCES INJUSTIFIEES

Dans ce paragraphe, nous essayerons de répondre à trois questions principales:

1. Quel type de lemme est inséré dans un dictionnaire papier, et quel type en est exclu?
2. Quelles normes sont appliquées pour inclure/exclure les lemmes?
3. Les théories et les méthodes de structuration d'un dictionnaire électronique peuvent-elles être utiles pour rédiger un dictionnaire papier?

Pour répondre à ces questions, il est nécessaire de passer au crible le domaine du dictionnaire papier, c'est-à-dire la lexicographie, que l'on peut ainsi définir:

“lessicografia

Tecnica di compilazione di dizionari; si avvale degli studi di lessicologia, semantica, morfologia e sintassi e ovviamente della metalessicografia, cioè degli studi teorici sul fare i dizionari, sul disegnare i lemmari e le voci secondo i bisogni degli utenti, le caratteristiche del tipo di dizionario e delle lingue descritte e usate nei dizionari. Per lessicografia si intende anche l'insieme di opere lessicografiche prodotte in una data lingua o in un dato periodo. (...) La linguistica computazionale ha migliorato le tecniche della lessicografia, e ha facilitato la compilazione di concordanze, liste di frequenza, dizionari inversi, rimari. La cosiddetta “corpus linguistics”, la linguistica che si basa su studi condotti a partire da corpora di testi su

supporto elettronico, ha facilitato la redazione di liste di parole nuove e diffuso in lessicografia l'abitudine a ricorrere a esempi tratti da corpora e ad organizzare lemmario e glosse tendendo conto della frequenza d'uso." ⁴⁴

Cette citation, non seulement définit un cadre général de référence pour le dictionnaire papier, mais spécifie aussi l'apport donné par la linguistique-informatique à la lexicographie en ce qui concerne les techniques de repérage et de consultation automatique des données. Grâce à la *corpus linguistics*, il est en fait possible de localiser de nouveaux mots, de calculer les fréquences d'usage et d'utiliser les résultats de l'analyse textuelle automatique pour l'élaboration de listes de lemmes et de gloses. La possibilité de vérifier la qualité des listes de lemmes déjà existantes n'est pas prise en considération, c'est-à-dire que l'on ne voit pas la nécessité de réviser la matière lexicale déjà décrite⁴⁵, opération qui permettrait de corriger des fautes, de circonscrire des lacunes descriptives, donc de légitimer qualitativement le lexique déjà répertorié.

En général, les maisons d'éditions effectuent des révisions continues et scrupuleuses de leurs dictionnaires, même si probablement ces révisions sont occasionnées plus fréquemment par des exigences typographiques – comme les corrections d'épreuves – que par la volonté d'effectuer une analyse lexicographique rétrospective. Néanmoins, nous ne devons pas oublier que le dictionnaire fait partie de la lexicographie mais n'est pas suffisant pour la définir, vu qu'il est un instrument lexicographique descriptif et non normatif. Un dictionnaire papier ratifie les usages linguistiques, mais il n'est pas à même de les créer ou de les imposer. Donc, l'approche critique envers le lexique, c'est-à-dire l'étude des néologismes et la révision des listes de

⁴⁴ **Lexicographie.** Technique de compilation de dictionnaires; elle se base sur les études de lexicologie, de sémantique, de morphologie et de syntaxe et naturellement de la méta-lexicographie, c'est-à-dire les études théoriques sur comment fabriquer les dictionnaires ou comment tracer les listes de lemmes selon les besoins des usagers, les caractéristiques du type de dictionnaire et des langues décrites et utilisées dans les dictionnaires. La lexicographie est aussi l'ensemble des œuvres lexicographiques produites dans une langue donnée ou pendant une période spécifique. (...) La linguistique-informatique a amélioré les techniques de concordances, de listes de fréquence, de dictionnaires inverses, de dictionnaires de rimes. La "corpus linguistics" ainsi dite, la linguistique qui se base sur des études menées à partir de corpus de textes sur support électronique, a facilité la rédaction de liste de nouveaux mots et a diffusé dans la lexicographie l'usage de se référer à des exemples pris dans des corpus et à organiser la liste de lemmes et les gloses en tenant compte de la fréquence d'usage." Beccaria G. L. ouvrage cité (c'est nous qui traduisons).

⁴⁵ Vu que nous avons présenté seulement un extrait de l'entrée *lessicografia*, nous précisons que ces aspects ne sont même pas pris en examen dans les parties que nous n'avons pas transcrites.

lemmes déjà existantes, est ou bien devrait être une tâche de la lexicologie, la discipline dont est dérivée la lexicographie. De la lexicologie nous donnons ici une brève définition, surtout pour vérifier la présence d'une telle approche critique:

“lessicologia

Lo studio dei lessemi di una lingua, delle loro relazioni, dei cambiamenti della loro forma e significato nel tempo. Poiché gli studi di lessicologia sono d'interesse per la lessicografia, si è anche affermata un'accezione ristretta di lessicologia come momento teorico della preparazione delle opere lessicografiche; quest'accezione viene però gradualmente abbandonata a favore dell'uso del termine metalessicografia (...) nell'ultimo quarto del sec. XX il termine lessicologia è stato spesso trascurato a favore di altri, quali morfologia, morfosintassi, semantica, semantica lessicale. Il fenomeno rispecchia una tendenza in fondo positiva e cioè la convinzione che il lessico sia molto più strutturato e molto più collegato alla sintassi di quanto non si volesse ammettere in passato (...) Ambiti classici e incontestati della lessicologia sono quelli che riguardano la formazione nel tempo del lessico di una lingua attraverso l'apporto di altre lingue (etimologia; calco; forestierismo) e la creazione di neologismi. Gli studi sulla formazione delle parole composte e derivate, sugli affissi sono anche un ambito tradizionale della lessicologia che attualmente viene sviluppato soprattutto in ricerche di morfosintassi, nel tentativo di cogliere meglio l'aspetto strutturato del lessico e i rapporti fra lessico e sintassi. La lessicologia è stata la prima area degli studi linguistici a beneficiare grandemente dei metodi quantitativi e dell'elaborazione elettronica di corpora linguistici (...)⁴⁶.

⁴⁶ **“Lexicologie.** L'étude des lexèmes d'une langue, de leurs relations, des changements de leur forme et de leur signifié dans le temps. Vu que les études de lexicologie sont importantes pour la lexicographie, une interprétation restreinte de la lexicologie s'est affirmée comme moment théorique de la préparation des œuvres lexicographiques; cette interprétation est toutefois graduellement abandonnée en faveur de l'usage du terme méta-lexicographie (...) pendant le dernier quart du XXème siècle le terme lexicologie a souvent été négligé en faveur d'autres termes, tels que morphologie, morphosyntaxe, sémantique, sémantique lexicale. Le phénomène reflète une tendance en tout cas positive, c'est-à-dire la certitude que le lexique est beaucoup plus structuré et beaucoup plus lié à la syntaxe que ce que l'on voulait admettre au temps passé (...) Les domaines classiques et incontestés de la lexicologie sont ceux qui s'occupent de la formation du lexique d'une langue dans le temps grâce à l'apport d'autres langues (étymologie, calque,

L'approche critique lexicologique se base donc sur une analyse du lexique ponctuelle et constante, synchronique et diachronique, qui a comme but aussi celui de cerner les éléments à décrire et les propriétés avec lesquelles effectuer les descriptions. Les dictionnaires papier sont seulement un reflet des études lexicologiques; et trouver les qualités et les défauts d'une œuvre lexicographique veut dire les trouver aussi et indirectement dans la lexicologie. Il est important de souligner que, dans la citation précédente, on exprime la *certitude que le lexique est beaucoup plus structuré et beaucoup plus lié à la syntaxe que ce que l'on voulait admettre au temps passé*; comme nous le verrons, ceci est un fondement du lexique-grammaire. En outre, on reconnaît aussi les bénéfices qui proviennent de la linguistique-informatique, en terme d'usage de l'ordinateur dans l'étude du langage naturel.

Ce que nous venons de citer nous permet aussi d'observer que la lexicologie, la lexicographie et la linguistique-informatique ont non seulement des objectifs communs, mais aussi échangent et s'empruntent des méthodologies d'investigation dans le but d'améliorer leurs méthodes d'enquête et les résultats que ces méthodes produisent. Donc, d'un point de vue du catalogage et de la description du lexique, linguistique-informatique d'un côté, lexicologie et lexicographie de l'autre, devraient aspirer à rejoindre les mêmes conclusions, d'un point de vue qualitatif et quantitatif.

Néanmoins, après une analyse attentive des dictionnaires papier, nous sommes obligés d'affirmer que cette correspondance de résultats est seulement hypothétique, parce que par rapport aux œuvres électroniques, les œuvres sur papier présentent de sérieux démaillages structuraux, que nous pouvons ainsi résumer:

barbarisme) et à la création de néologismes. Les études sur la formation des mots composés et des dérivés et sur les affixes sont aussi des domaines traditionnels de la lexicologie, qui sont développés actuellement surtout dans les recherches de morphosyntaxe, dans le but de mieux saisir l'aspect structuré du lexique et les rapports entre lexique et syntaxe. La lexicologie a été le premier terrain des études linguistiques qui ait considérablement profité des méthodes quantitatives et de l'élaboration électronique des corpus linguistiques (...)“. Beccaria, G. L. ouvrage cité (c'est nous qui traduisons).

- listes de lemmes incomplètes, avec des descriptions qui souvent varient beaucoup d'un dictionnaire à l'autre d'une maison d'édition, comme si chaque dictionnaire décrivait une matière différente et indépendante;
- souvent, aucun ensemble de normes apparemment utiles pour justifier la lemmatisation ou la non-lemmatisation de mots spécifiques;
- là où les normes sont au contraire annoncées à l'avance et expliquées, leur application non constante est un facteur qui en réduit la valeur et la légitimité.

Ce que nous venons d'observer ne provient pas seulement de la comparaison que nous avons faite entre dictionnaires papier et électronique, mais est aussi matière à discussion à l'intérieur de la même lexicographie. Pour démontrer cela, nous présentons un passage dans lequel le système lexicographique du français est profondément critiqué, même s'il faut souligner que le système français est en quelque mesure plus complet et soigné que l'italien:

"Affirmer que les dictionnaires de mêmes dimensions sont fondamentalement identiques, moyennant certaines différences de détail, est une banalité courante chez les lecteurs et souvent chez les linguistes, et pourtant cette affirmation est radicalement fausse. Cette erreur relève d'une méconnaissance de la lexicographie, mais aussi du français qui y est décrit. Ces dictionnaires sont au contraire totalement différents:

- *sur la nature et le nombre des entrées;*
- *sur les nomenclatures;*

- sur la structure des articles;
- sur les informations données;
- sur le français décrit.

*(...) l'image que les dictionnaires donnent de la langue est très différente; c'est la conséquence de divergences linguistiques, culturelles, littéraires, et de différences dans les objectifs visés.(...) l'objectif final est soit l'usage linguistique du mot, soit la description de l'objet ou de l'être désigné (...) le vocabulaire dit général comporte tous les termes qui ont pu être relevés dans les communications quotidiennes et, en particulier, dans les textes écrits (journaux ou littérature) (...) traditionnellement, le mot introduit dans un dictionnaire doit avoir été relevé, attesté dans plusieurs textes. Mais cette notion d'attestation ne peut pas plus avoir la même valeur dans une société où les médias sont à la fois plus nombreux et plus divers, et où la quantité d'information et la variété des communications sont incommensurables au sens premier du terme."*⁴⁷

Dans notre analyse de ces arguments, nous avons pu observer que les notations faites à propos du système lexicographique français sont valables et pertinentes aussi pour le système italien, et pour le prouver nous avons fait un essai rapide, en sélectionnant quelques tranches lexicales de l'italien et en recherchant dans les différents dictionnaires les mots qu'ils incluent. Les tranches lexicales que nous avons choisies sont:

- de *abbattibile* (abattable) à *abbattuto* (abattu);
- de *costruibile* (constructible) à *costruzione* (construction);

⁴⁷ Dubois, J., Dubois-Charlier, F., "Incomparabilité des dictionnaires", dans Courtois, B., Silberztein, M., 1990, page 5.

- de *struttura* (structure) à *strutturistico* (structuristique).

La lemmatisation des mots inclus dans ces tranches lexicales a été contrôlée aussi, outre que sur le DELAS, sur les dictionnaires électroniques et informatisés suivants:

- *Il dizionario della Lingua Italiana su CD-Rom* de Giacomo Devoto et Giancarlo Oli;
- *Lo Zingarelli 2002 in CD-Rom*, de Nicola Zingarelli;
- le site Web www.garzanti.it et le moteur de recherche *Digita-Web, Il Dizionario Garzanti della Lingua Italiana*.

Dans la table suivante, nous donnons un résultat comparatif de la recherche effectuée:

	DELAS	Devoto - Oli	Zingarelli 2002	Garzanti
1.	abbattibile	abbattibile	abbattibile	-
2.	abbattibilità	-	-	-
3.	abbattifieno	abbattifieno	-	abbattifieno
4.	abbattimento	abbattimento	abbattimento	abbattimento
5.	abbattitore	abbattitore	abbattitore	abbattitore
6.	abbattuta	abbattuta	abbattuta	abbattuta
7.	abbattutamente	-	-	-
8.	abbattuto	abbattuto	abbattuto	abbattuto
9.	costruibile	costruibile	costruibile	costruibile
10.	costruibilità	-	-	-
11.	costruibilmente	-	-	-
12.	costruire	costruire	costruire	costruire
13.	costruito	costruito	-	-
14.	costruttivamente	-	costruttivamente	costruttivamente
15.	costruttivismo	costruttivismo	costruttivismo	costruttivismo
16.	costruttivista	-	-	-
17.	costruttivisticamente	-	-	-
18.	costruttivistico	-	-	-
19.	costruttivo	costruttivo	costruttivo	costruttivo
20.	costrutto	costrutto	costrutto	costrutto
21.	costruttore	costruttore	costruttore	costruttore
22.	costruzione	costruzione	costruzione	costruzione
23.	struttura	struttura	struttura	struttura
24.	strutturabile	strutturabile	strutturabile	strutturabile
25.	strutturabilità	-	-	-
26.	strutturabilmente	-	-	-
27.	strutturale	strutturale	strutturale	strutturale
28.	strutturalismo	strutturalismo	strutturalismo	strutturalismo
29.	strutturalista	strutturalista	strutturalista	strutturalista
30.	strutturalisticamente	-	strutturalisticamente	-
31.	strutturalistico	strutturalistico	strutturalistico	-
32.	strutturalità	-	-	-
33.	strutturalmente	-	strutturalmente	-
34.	strutturante	strutturante	-	-
35.	strutturare	strutturare	strutturare	-
36.	strutturatamente	-	-	
37.	strutturato	strutturato	strutturato	strutturato
38.	strutturazione	strutturazione	strutturazione	strutturazione
39.	strutturista	strutturista	strutturista	strutturista
40.	strutturistica	strutturistica	strutturistica	-
41.	strutturistico	-	-	-

Table 6

Dans la table ci-dessus, nous avons numéroté les lemmes inclus dans le DELAS et les avons mis à côté de ceux que nous avons trouvé dans les dictionnaires papier, en

signalant avec des lignes les correspondances manquées, pour aider à la comparaison entre les différentes listes. Il est donc aisé de noter que, par rapport au DELAS, les lemmes dans le dictionnaires papier sont en nombre inférieur, et nous observons aussi que la présence dans le DELAS d'un nombre plus élevé de lemmes n'est pas due à une surproduction arbitraire d'unités lexicales, mais elle est produite par l'insertion de mots dont l'existence, la compréhension et la cohérence ont été raisonnablement évaluées avant d'en approuver l'inclusion dans la liste de lemmes. En fait, nous savons que les lemmatisations du dictionnaire électronique sont réglées par une série de normes spécifiques, parmi lesquelles nous trouvons aussi des tests à effectuer avec les locuteurs, et en dernière analyse, par l'expérience des linguistes. Sur toutes ces bases, il est possible d'affirmer que la taille plus grande d'un dictionnaire électronique, comparée à celle d'un dictionnaire papier, n'est pas due à l'ajout de lemmes inexistants, créés artificiellement et sans signifié précis, mais à la lemmatisation de mots avec des signifiés bien définis, donc facilement utilisés et utilisables dans la langue italienne.

Les listes de la table (6) mettent aussi en évidence les différents choix de lemmatisation appliqués par les quatre dictionnaires, et aussi l'existence de véritables "vides lexicaux" à l'intérieur des dictionnaires papier, mais elles ne nous donnent aucun élément utile pour individualiser les critères grâce auxquels le *Garzanti* a lemmatisé *costruibile* et non *abbattibile*, ni elles ne nous font comprendre pourquoi le *Devoto-Oli* n'a lemmatisé aucun adverbe en *-mente*⁴⁸.

En outre, ces vides lexicaux se révèlent encore plus singuliers si l'on considère ce que l'on a déjà dit à propos des nouvelles tendances de la lexicologie et de la lexicographie, surtout en ce qui concerne l'apport fourni à ces disciplines par la linguistique-informatique qui devrait rendre plus aisé le dépouillage de corpus textuels même très vastes et par conséquent le repérage des nouveaux mots avec lesquels mettre à jour les listes de lemmes.

⁴⁸ On pourrait répliquer que beaucoup d'adverbes dans le *Devoto-Oli* sont indiqués dans les descriptions faites pour les adjectifs dont ils dérivent, et que donc *costruttivamente* (constructivement) pourrait être signalé en relation avec *costruttivo* (constructif). En tout cas, ce type de notation ne nous ne permet pas d'affirmer que les adverbes du *Devoto-Oli* ont une lemmatisation indépendante, mais seulement qu'ils sont indiqués comme des entrées ou des usages dérivés de certains adjectifs.

En comparant tous les mots du DELAS avec ceux d'autres ouvrages, il a été possible de définir assez précisément quelles typologies de mots sont insuffisamment listées dans les dictionnaires papier qui présentent des absences injustifiées mais non systématiques. En fait, les lemmes qui appartiennent à certains groupes peuvent apparaître dans un dictionnaire et être absents dans un autre. En même temps, dans un même dictionnaire, les éléments d'un groupe donné peuvent être lemmatisés irrégulièrement, c'est-à-dire que l'on en trouvera certains tandis que d'autres seront absents. Les notations qui suivent concernent beaucoup de dictionnaires papier italiens, y compris les différentes versions du *Zingarelli* et du *Devoto-Oli*. Nous résumons schématiquement les typologies de lemmes absents:

- les participes passés à usage adjectival, comme par exemple *costruito* (bâti) dans la phrase *questo edificio è costruito bene* (cet édifice est bien bâti);
- les participes présents à usage adjectival, comme par exemple *amareggiante* (attristant) dans la phrase *questa situazione è amareggiante* (cette situation est attristante);
- les adjectifs dénominaux, comme par exemple *viscometrico* (viscosimétrique) dérivé de *viscometria* (viscosimétrie);
- les adverbes en *-mente*, comme par exemple *allucinatoriamente* (hallucinatoirement);
- les verbes en *-izzare*, comme par exemple *routinizzare* (routiniser), et par conséquent tous les participes présents et passés à usage adjectival dérivés de ces

verbes, comme *routinizzante* (routinisant) et *routinizzato* (routinisé), et aussi les déverbaux en *-zione* comme par exemple *routinizzazione* (routinisation);

- les verbes pronominaux dérivés d'autres verbes transitifs ou intransitifs. A ce sujet, il faut souligner que dans certains cas les verbes pronominaux sont signalés seulement comme des emplois des verbes dont ils dérivent, ce qui se vérifie avec *abbassarsi* (s'abaisser) et *scomodarsi* (se déranger), tandis que, dans d'autres cas, des formes à haute occurrence, comme *aspergersi* (s'asperger) et *rimboccarsi* (se retrousser) ne sont pas présentes dans les listes de lemmes;
- les adjectifs en *-bile*, comme par exemple *attenuabile* (atténuable);
- les substantifs en *-bilità*, comme par exemple *attenuabilità* (la possibilité d'être atténué);
- les lemmes obtenus grâce à une suffixation, comme par exemple *acomunista* (non-communiste), *acomunistico* (non-communistique), *reincollare* (recoller), *ricollegabile* (branchable à nouveau), *rimescolatura* (re-mélange) et *scollegabile* (débranchable);
- l'indication de l'usage adjectival pour les substantifs en *-tore* ou *-one*, comme par exemple *assolutore* (celui qui donne l'absolution) et *cialtrone* (canaille), comme dans *sentenza assolutrice* (sentence d'absolution) et *politica cialtrona* (politique de canaille);

- l'étiquette de *substantif* pour plusieurs adjectifs, non définissables en terme de classe suffixale, mais qui incluent des mots comme *gli accampati* (personnes qui sont campées), *gli alternativi* (alternatifs), *gli assunti* (gens engagés), *i coraggiosi* (les courageux), *i lussuriosi* (les luxurieux) et ainsi de suite;
- la flexion au féminin de quelques substantifs à référents humains, comme par exemple *campionissimo* (le champion le plus grand), *dispersore* (disperseur, et en italien *disperditrice*⁴⁹ au féminin) ou *inquisitore* (inquisiteur, et en italien avec le féminin *inquisitrice*, si utilisé dans le sens non historique du terme);
- l'étiquette de substantif seulement au singulier pour quelques dialecte comme par exemple *barese* (dialecte de la ville de Bari), *cagliaritano* (dialecte de la ville de Cagliari) ou *lucchese* (dialecte de la ville de Lucca);
- les noms propres seulement au singulier relatifs à des zones géographiques spécifiques, i.e. les alentours des grandes villes, comme par exemple la *Lucchesia* (les alentours de la ville de Lucca), le *Comasco* (les alentours de la ville de Como), le *Palermitano* (les alentours de la ville de Palermo);

⁴⁹ À part quelque cas, les unités lexicales italiennes qui terminent au masculin en *-ore* sont soit des adjectifs soit des substantifs à référent humain, et donc ils ont au moins l'une des formes féminines au singulier en *-rice*. Ces unités lexicales sont des déverbaux et sont presque toujours formées avec l'adjonction du suffixe *-ore* à la racine du participe passé du verbe correspondant, comme par exemple *istigat-ore* (instigateur, à partir de *istigato*, instigué, participe passé du verbe *istigare*, instiguer). Néanmoins, cette règle n'est pas valable pour un nombre consistant de mots, pour lesquels la formation du féminin est faite à partir de la racine de l'infinitif et qui modifient aussi le suffixe, comme par exemple *dissuad-itrice*, féminin de *dissuasore* (qui dissuade). Dans ces formes, le suffixe *-rice* est souvent précédé par la séquence *-it-*, comme il est possible de le vérifier sur d'autres mots tels que *aggre-ditrice* (féminin de *aggressore*, aggresseur), *aspergitrice* (féminin de *aspersore*, asperseur), *difenditrice* (féminin de *difensore*, défenseur), *interceditrice* (féminin de *intercessore*, intercesseur) et ainsi de suite. Il n'est pas aisé de motiver ces formations particulières, mais on peut affirmer qu'elles ne sont pas toutes mentionnées dans les principaux dictionnaires papier italiens.

- les noms de quelques ères géologiques, comme par exemple *Aaleniano* (Aalénien) ou *Burdigaliano* (Burdigalien), et aussi les formes adjectivales respectives utilisées dans les expressions *il periodo aaleniano* (la période aalénienne), *l'era burdigaliana* (l'ère burdigalienne);
- la flexion non-régulière au féminin des substantifs à référent humain, comme par exemple *evangelista* (évangéliste) et *mormone* (mormon).

Ce dernier point mérite une analyse plus approfondie, parce qu'il a une portée majeure par rapport à ceux qui le précèdent et aussi parce qu'il vise la description morphologique effectuée dans les dictionnaires papier à propos des substantifs non épïcènes, qui différencient formellement leur flexion masculine de la flexion féminine. La description de ces termes paraît imprécise et non systématique, surtout en rapport à des domaines lexicaux spécifiques tels que les noms de profession, des membres d'associations religieuses et sportives, de courants littéraires, philosophiques et artistiques. Pour ces noms, on ne signale pas toujours la correspondance entre genre et forme, c'est-à-dire la correspondance qui existe entre un référent tel que "femme qui exerce la profession d'avocat" et le mot *avvocatessa* (avocate) qui sert à l'indiquer. Par exemple, dans *Lo Zingarelli 2001* le lemme *mormone*, qui est utilisé en référence aux membres de la communauté religieuse, est lemmatisé seulement comme substantif masculin, ce qui exclut donc l'occurrence féminine possible *una mormone* (une mormone). Toutefois, dans le même dictionnaire, pour le substantif *quacchero* (quaker) on trouve l'indication d'une flexion féminine en *quacchera*.

Les deux termes précédents sont sémantiquement contigus, parce qu'ils indiquent deux communautés religieuses insérées dans des contextes culturels et sociaux plutôt similaires, comme par exemple ceux de l'Amérique du Nord. Ces deux mots peuvent donc être emblématiques pour notre analyse, soit en ce qui concerne la classification morphologique adoptée dans les dictionnaires papier, soit pour les conséquences d'une lemmatisation éventuellement erronée. Des deux substantifs, qui

sont tous deux des emprunts de l'anglais d'Amérique, l'italien a reçu surtout l'usage figuré, même si par rapport à *quacchero*, *mormone* est aussi utilisé dans une acception plus littéraire. Néanmoins, une description morphologique incomplète de ce dernier mot, et en termes plus vastes de tous les mots non épicènes, dénote d'un côté une analyse non ponctuelle des unités sémantiquement contiguës et qui appartiennent à un même groupe lexical – celui relatif aux noms des communautés religieuses – tandis que d'un autre côté elle impose à l'usage linguistique des limites arbitraires, non réels parce que non dérivés de facteurs pratiques ou empiriques. Comme nous le verrons, de telles fautes dans la lemmatisation, qu'elle soit absente, incomplète ou relative à des mots simples, peuvent produire de sérieuses chutes culturelles.

Il y a d'autres typologies mineures de lemmatisation absente ou incomplète, représentées par des exemples simples et appartenant à presque toutes les catégories grammaticales. Puisqu'il est difficile de les résumer schématiquement, nous en citons quelques cas:

- le verbe *imbranarsi* (s'entraver);
- les adjectifs *ex*, comme dans la phrase *la mia ex fidanzata* (mon ex-fiancée) et *quasi* (presque) comme dans *un mio quasi amico* (un de mes presque copains);
- les substantifs féminins pluriels relatifs aux heures de la journée, comme *le tredici* (treize heures) et *le ventiquattro* (vingt-quatre heures);
- l'usage pronominal des numéraux et des cardinaux, comme dans les phrases *ne ho visto uno* (j'en ai vu un) et *sono arrivato primo* (je suis arrivé premier).

La description du lexique italien, dans son état actuel, est donc fragmentaire et fragmentée, pourtant le fait d'avoir localisé cette fragmentation ne nous aide malheureusement pas à en déterminer les causes. Dubois et Dubois-Charlier⁵⁰ affirment qu'insister sur les différences qui existent entre les dictionnaires papier et aussi sur leurs lacunes a un sens seulement par rapport à un usage utopique de la langue, où tout est clair et bien défini. Les dictionnaires ne font pas la description d'une langue idéale, mais ils correspondent seulement à l'idée que le lexicographe s'est faite de la langue et dépendent du progrès des sciences, et en particulier de la linguistique. Si chaque dictionnaire représente une tentative de description particulière d'un objet, il faut rappeler qu'il ne peut être confondu avec l'objet décrit: cette confusion est néanmoins à la base de l'opinion typique et parfois mystificatrice que l'on a des dictionnaires papier, selon laquelle les descriptions qu'ils adoptent sont les seules possibles.

Selon Dugas⁵¹, la compilation des entrées lexicales dans un dictionnaire papier répond surtout à des critères de fréquence d'usage, et les différents milliers de mots qu'ils contiennent sont largement suffisants à couvrir les besoins communicatifs des locuteurs. Les limites de l'élaboration des dictionnaires actuels dépendent de problèmes de documentation, de réaction et d'édition, du respect d'un concept de norme qui comporte le rejet de mots considérés offensifs, obsolètes, trop techniques et ainsi de suite. Néanmoins, les lexicologues doivent admettre que dans les dictionnaires papier on ne dit pas tout à propos des propriétés lexicales et de leur usage dans les règles productives de la formation des mots. Donc, il sera nécessaire d'ajouter un grand nombre d'entrées dans les dictionnaires papier qui sont dans le commerce, et aussi revoir les entrées que des dictionnaires contiennent déjà pour que ces œuvres puissent rejoindre une couverture utile pour décrire une langue de façon réaliste.

Pour aller de pair avec le progrès des sciences, et en particulier de la linguistique, un dictionnaire électronique doit donc être toujours en phase de mise à jour, parce que malgré sa taille déjà remarquable, pour les raisons que nous avons indiquées, il ne peut pas inclure tous les mots utilisables dans une langue. Mais il vaut la

⁵⁰ Dubois, J., Dubois-Charlier, F. ouvrage cité.

⁵¹ Dugas, A., "La création lexicale et les dictionnaires électroniques", dans Courtois, B., Silberztein, M., ouvrage cité.

peine de rappeler que, comme outil d'analyse textuelle automatique, le dictionnaire électronique observe la langue d'usage, en donnant de celle-ci un compte-rendu sûrement plus réaliste que celui donné par les dictionnaires papier. En ce sens, l'efficacité majeure du dictionnaire électronique ne représente pas seulement un aspect quantitatif, mais aussi qualitatif si nous considérons la nature des informations associées aux mots simples ou aux mots composés.

Sur la base de ce que nous venons d'observer, nous pouvons dire que les lacunes typiques des dictionnaires papier dérivent directement d'une approche imparfaite de l'étude du lexique, en ce qui concerne le catalogage ou la description. On peut seulement supposer que la cause principale en soit le dépouillement de corpus pas suffisamment vastes, qui sont les seuls permettant une mise à jour exhaustive des listes de lemmes, en terme de quantité des unités lexicales et d'exemplification de leurs contextes distributionnels possibles. Ce type d'approche semble en outre plutôt généralisée, vu qu'elle se vérifie non seulement avec les dictionnaires papier, mais aussi avec d'autres œuvres qui traitent du langage naturel, telles que, par exemple, les grammaires et les dictionnaires grammaticaux, qui devrait au contraire être les lieux idéaux pour une étude exhaustive des aptitudes humaines à la verbalité. Il semble donc nécessaire de modifier l'approche des corpus, en étendant le choix et les modalités d'étude, c'est-à-dire en analysant des corpus toujours plus vastes et typologiquement hétérogènes. Il ne reste qu'à évaluer quels avantages peuvent dériver d'un tel élargissement de l'analyse. Puisque le problème de la mise à jour de la liste des lemmes, dans les prochains paragraphes nous donnerons les résultats de deux analyses que nous avons effectuées en dépouillant des corpus certes marginaux mais très riches en ce qui concerne les mots aussi non-lemmatisés dans notre dictionnaire électronique.

3.1.1 L'affaire Perniola

Le premier corpus que nous avons analysé est le livre *Il Sex-appeal dell'inorganico* (Le Sex-appeal de l'inorganique) du philosophe italien Mario Perniola (1994), œuvre qui a l'avantage d'être à la fois lexicalement innovatrice et profondément enracinée dans les pratiques et les méthodes de la critique philosophique. Les thèmes centraux du livre sont la virtualité et l'inorganicité qui se manifestent aujourd'hui à l'intérieur des relations interpersonnelles. Perniola traite de ces arguments en utilisant les mots du "vieux" lexique de la philosophie, sans néanmoins renoncer à un usage créatif de la langue, orthodoxement basé sur les normes de la morphologie dérivationnelle de l'italien.

Nous avons voulu vérifier quels mots utilisés par Perniola ne figuraient pas dans les dictionnaires papier et dans notre dictionnaire électronique. Nous avons ainsi créé la liste suivante de nouveaux mots dans laquelle aux lemmes simples nous ajoutons les étiquettes des catégories grammaticales d'usage. Dans quelques cas, pour être plus précis, nous avons ajouté à des substantifs l'étiquette *solo sing.* (seulement singulier)⁵²:

⁵² Par exemple, le mot *antropologico, N (solo sing.)* (l'anthropologique) a la même valeur grammaticale que *il ridicolo* (le ridicule) dans des expressions du type *scadere nel ridicolo* (tomber dans le ridicule) et *parlare dell'antopologico* (parler de l'anthropologique).

accessoriamente, ADV
 adelfico, A
 afanisi, N
 anonimità, N
 antidialettico, A
 antimoda, N
 antropologico, N (solo sing.)
 antropologizzare, V RSI
 antropologizzato, A
 antropologizzazione, N
 antropomorfizzare, V
 antropomorfizzato, A
 antropomorfizzazione, N
 artificiale, N (solo sing.)
 autoannientamento, N
 autoannientarsi, V
 autocosciente, A
 autodisciplinarsi, V
 autoevidente, A
 autoevidenza, N
 autonomizzare, V
 autonomizzato, A
 autonomizzazione, N
 autonumerarsi, V
 autopenetrarsi, V
 autopenetrazione, N
 autorappresentazione, N
 autoreferenziale, A
 autoriferimento, N
 autoriflessione, N
 autoriflessivo, A
 autorispecchiarsi, V
 autosoppressione, N
 autosopprimersi, V
 autosuperamento, N
 coappartenenza, N
 concupito, N (solo sing.)
 cosalità, N
 cromatimico, A
 cunnilinctus, N
 cybersex, N
 defunzionalizzare, V
 defunzionalizzato, A
 defunzionalizzazione, N
 desoggettivizzazione, N
 desoggettivare, V
 desoggettivato, A
 deumanizzante, A
 deumanizzare, V
 deumanizzato, A
 deumanizzazione, N
 disincarnato, A
 disponibile, N (solo sing.)
 divorante, A
 domestico, N (solo sing.)
 eiaculante, A
 enantiodromia, N
 erranza, N
 esistente, N (solo sing.)
 eternità, N
 eteroreferente, A
 eteroreferenziale, A
 fattità, N
 funerario, N (solo sing.)

geotopico, A
 heideggeriano, A
 heideggeriano, N
 hopi, A
 hopi, N
 idiogamia, N
 idiogamico, A
 idiomania, N
 imbestialimento, N
 impartecipe, A
 impensabile, N (solo sing.)
 impensato, N (solo sing.)
 inabitabile, N (solo sing.)
 inaccessibile, N (solo sing.)
 incondizionatezza, N
 indistinzione, N
 indumentale, A
 inessenziale, A
 inessenzialità, N
 irraggiungibile, N (solo sing.)
 mereologia, N
 mereologico, A
 metaletterario, A
 metaletteratura, N
 metanarrativo, A
 metanarrazione, N
 metascrittura, N
 metasessuale, A
 metasessualità, N
 micidialità, N
 mondanizzare, V
 mondanizzato, A
 mondanizzazione, N
 naturale, N
 neostoicismo, N
 neostoico, A
 neostoico, N
 noumenico, A
 oggettività, N
 oggettivante, A
 omeostasi, N
 ontologizzante, A
 ontologizzare, V
 ontologizzato, A
 ontologizzazione, N
 organico, N
 orgasmolatria, N
 orgasmomania, N
 originario, N (solo sing.)
 parasportivo, A
 penetrante, N
 penetrato, N
 pensabile, N (solo sing.)
 percepiente, N
 pigmalionismo, N
 polverosamente, ADV
 postumano, A
 postvitale, A
 prefunerario, A
 premortuario, A
 presentificazione, N
 preumano, A
 progettante, A
 programmante, A

protosessuale, A
 protosessualità, N
 prototetica, N
 prototetico, A
 pseudopode, N
 rappacificante, A
 reificante, A
 reificatorio, A
 schellinghiano, A
 schellinghiano, N
 sensoristica, N
 sensoristico, A
 sentito, N (solo sing.)
 senziante, N
 sessualismo, N
 socializzante, A
 soddisfabile, A
 sopprimibilità, N
 sotterranemante, ADV
 sovraelevazione, N
 spermafogo, A
 spiritualizzante, A
 splatterpunk, N
 statuofilia, N
 tecnologico, N (solo sing.)
 tecnomorfismo, N
 testimoniante, A
 transfinito, A
 urbano, N
 vampirizzato, N
 vestimentale, A
 villagiano, N
 visivo, N
 wittgensteiniano, A

Les nouveaux lemmes repérés, pris d'un texte qui, en termes informatiques, pourrait avoir une taille de presque 600 kilobytes, sont au nombre de 163. Ces mots ne sont pas répertoriés dans la plus part des dictionnaires papier italiens, même s'il s'agit de lemmes non spécifiquement philosophiques, comme par exemple *metanarrativo* (méta-narratif), *metanarrazione* (méta-narration) ou *rappacificante* (réconciliant). Si nous considérons que le domaine de connaissance concerné est spécialisé, et donc "marginal" en termes de diffusion, l'apport est certainement très important et il est licite de penser que d'autres textes du même genre peuvent contenir un nombre aussi élevé de nouveaux lemmes⁵³. Le sens des mots est surprenant: ils sont tous suffisamment transparents du point de vue de la sémantique et de morphologie dérivationnelle. Ils sont aussi largement utilisables.

Après l'analyse de l'ouvrage de Mario Perniola, ces 163 nouveaux mots ont été listés et fléchis dans le DELAS-DELAFF de l'italien.

⁵³ Voir aussi AA. VV., 1988, dont ne nous avons analysé que quelques entrées.

3.1.2 Freud versus Popper, i.e. la Psychanalyse versus le Faillibilisme

Nous choisissons, parmi d'autres, des exemples particuliers qui nous fascinent, tout en sachant qu'il faut tenir compte du fait qu'au départ il y a l'œuvre de médiation de la traduction. Nous estimons que les langues de spécialité dont nous parlons, celle de la psychanalyse surtout, ont une place dans la culture moderne, donc aussi dans la langue en général. Nous avons considéré leur impact sur l'italien, mais la même analyse peut être appliquée à d'autres langues.

Comme deuxième source de repérage lexical, nous avons utilisé les œuvres plus importantes de l'épistémologue Karl R. Popper, personnalité remarquable du XX^{ème} siècle qui a laissé un vaste héritage intellectuel dans le domaine de la philosophie et dans d'autres secteurs de la culture. En partant de l'indubitable valeur de sa pensée, nous avons analysé l'impact lexical de ses écrits traduits en italien, et pour avoir des critères objectifs de validation de cette analyse, nous en avons fait une similaire pour Sigmund Freud, autre personnalité extrêmement importante non seulement pour le dernier siècle, mais aussi pour ceux à venir. Comme nous le verrons, malgré l'énorme importance de ces deux personnages de la culture mondiale, les dictionnaires papier ont réservé leur des traitements très différents.

En 1866, Sigmund Freud ouvre à Vienne son cabinet privé pour la thérapeutique des maladies nerveuses. En se basant sur les observations de Joseph Breuer et en recherchant les raisons et les sens des manifestations hystériques, Freud entreprit l'usage de l'hypnose comme méthode curative ou de diagnostic, faisant les premières avancées vers la découverte de l'inconscient et donc vers la définition de la psychanalyse. Cette dernière, comme il est notoire, est devenue par la suite une thérapie à l'efficacité indiscutable, et aussi une clé de lecture de la vie moderne, dont il semble aujourd'hui que presque personne ne puisse faire abstraction.

Comme confirmer sa vocation à être la patrie de l'*intelligenza* européenne de cette période, en 1902 Vienne vit aussi naître Karl Raimund Popper, auquel on doit un important changement méthodologique dans l'approche épistémologique de la recherche scientifique. Popper, qui était un philosophe des sciences, constata que l'histoire de

toutes les théories scientifiques est articulée en trois phases essentielles, naissance, période d'affirmation maximale, déclin. En particulier, Popper affirme que le déclin de chaque théorie est dû à l'affirmation d'une autre théorie, plus innovatrice, qui revoit et corrige la précédente et qui donc, dans un premier moment la falsifie et ensuite la perfectionne. Dans ce cas, le philosophe autrichien reprend son idée de Hume, qui a démontré qu'il n'est pas possible d'affirmer la validité d'une théorie en généralisant la valeur d'un nombre fini d'événements qui la confirment ou qui se conforment à elle. En outre, selon Popper, il suffit d'avoir un seul événement contraire à ce qui a été prédiqué par la théorie pour l'invalider et avec elle invalider toutes les lois et les descriptions qu'elle inclut. Le grand changement méthodologique provoqué par la pensée de Popper est donc le suivant: on ne doit pas seulement postuler une théorie sur la base des événements qu'elle observe et explique, mais on doit aussi utiliser tous les événements observables comme des contrôles possibles à effectuer sur la théorie même, pour vérifier son efficace ou sa fausseté. Si un seul résultat des observations falsifie la théorie, alors elle doit être abandonnée, si au contraire toutes les observations la confirment, elle peut être considérée valable jusqu'au moment où elle ne sera pas inévitablement falsifiée par les contrôles futurs. Il s'ensuit que plus un système descriptif théorique est falsifiable, plus il peut être contrôlé par l'homme: plus la science se trompe, plus elle est à la portée de l'homme.

Donc, si aucune théorie scientifique n'est valide dans l'absolu, alors toutes les théories peuvent être comparées à de simples hypothèses ou conjectures sur le monde réel qu'elles étudient, parce qu'elles décrivent partiellement l'état des faits et peuvent être démenties par de nouveaux événements dans n'importe quel moment.

Pour support de ses thèses, à titre d'exemple, Popper cite les grandes étapes du progrès scientifique qui ont concerné l'astronomie, la physique et les sciences naturelles. Par exemple, les descriptions et les calculs de Galilée ont été valables jusqu'au moment où Newton les a revus et corrigés sur la base des nouveaux événements qu'il avait pu observer. A leur tour, les lois de Newton ont été soumises à un même procédé de révision et de correction par Einstein. Néanmoins, on ne peut pas affirmer que les calculs de Galilée et Newton étaient erronés, mais seulement qu'ils étaient partiellement vrais, parce qu'ils rendaient compte seulement d'une partie des événements observés. Au

contraire, les théories d'Enstein sont encore les meilleurs possibles parce qu'elles se rapprochent le plus de la description de la réalité, et aussi parce qu'elles n'ont pas encore été falsifiées par d'autres théories plus précises. L'idée de progrès qui sort de la théorie poppérienne est donc celle d'un lent et constant rapprochement d'une connaissance objective de la réalité, qui reste le but final de la science.

L'ensemble théorique et méthodologique conçu par Popper est dit faillibilisme, et encore aujourd'hui il suscite des débats plutôt vifs au sein des communautés philosophiques et scientifiques. Sa portée, comme on le notera, ne concerne pas seulement les théories scientifiques *stricto sensu*, mais, en qualité de doctrine philosophique elle a sans doute aussi une grande valeur en termes d'approche cognitive de la connaissance de la réalité, dans le sens plus vaste que cette expression peut avoir. Popper se fait l'écho en philosophie du scepticisme cognitif et de la désagrégation du monde observable qui ont distingués les messages d'autres secteurs culturels du XXème siècle: il suffit de penser à James Joyce, Virginia Woolf, Samuel Beckett, Henry Bergson, à Luigi Pirandello et au même Sigmund Freud, penché vers l'étude et la subdivision de la personnalité humaine qui semblait précédemment indivisible, comme indivisible semblait l'atome.

En termes d'histoire de la culture du monde, Freud et Popper ont eu une influence extraordinaire pendant la même période, c'est-à-dire le XXème siècle, tandis que du point de vue typiquement scientifique, psychanalyse et faillibilisme représentent deux des plus importants ensembles épistémologiques de toutes les époques. D'un autre côté, pour ce qui concerne plus précisément les arguments que nous traitons ici, nous dirons qu'avant 1866, c'est-à-dire avant que Freud ouvre son cabinet privé pour le traitement des maladies nerveuses, le mot *ipnosi* (hypnose) avait probablement des connotations différentes de celles d'aujourd'hui, et nous pourrions aussi douter du fait que ce mot était déjà utilisé dans la langue courante. Au contraire, des mots simples tels que *agorafobia* (agoraphobie), *nevrastenico* (neurasthénique), *isterico* (hystérique) ou des mots composés tels que *paura morbosa dei pulcini* (peur morbide des poussins), *allucinazione acustica* (hallucination acoustique) ou *mania di grandezza* (folie de grandeur) n'existaient sûrement pas. En fait, tous ces mots font partie du corpus technique que la psychanalyse utilise en qualité de discipline scientifique, et dans une

large mesure ils ont tous été créés et adoptés dans ce domaine, pour devenir ensuite partie intégrante du patrimoine lexical de chaque langue. Il est important de souligner que quelques-uns de ces termes, comme par exemple *isterico* ou *megalomane* (mégalomane), sont aujourd'hui aussi très utilisés dans des contextes non techniques et plutôt informels, ce qui confirme que la psychanalyse et sa terminologie ont des bases d'utilisation amples et bien fondées.

Nous n'avons pas trouvé un dictionnaire italien des termes de la psychanalyse, et donc il est difficile pour nous de calculer avec précision quel a été l'apport freudien au lexique de notre langue. Il est néanmoins possible d'affirmer que, sans cet apport, cette langue serait sûrement plus pauvre. Et si les termes de la psychanalyse sont amplement lemmatisés dans les dictionnaires papier, au contraire Popper et le lexique qu'il a forgé ou seulement utilisé avec des sens et des signifiés nouveaux ont reçu un autre type de traitement. Nous parlons donc de néologismes comme ceux qui sont insérés dans la liste suivante et qui ont été repérés dans des différentes œuvres de Popper, traduites en italien:

falsificabilmente (de façon falsifiable)
falsificazionismo (falsificationisme)
falsificazionista (falsificationiste)
falsificazionistico (de type falsificationiste)
fallibilismo (faillibilisme)
fallibilista (fallibiliste)
fallibilistico (de type fallibiliste)
popperiano (popperien)
popperianamente (de façon popperienne)

À ces néologismes, on doit aussi ajouter les usages particuliers et poppériens des mots tels que *falsificazione*, (falsification) *falsificabile* (falsifiable), *falsificabilità* (falsifiabilité).

Dans ce cas aussi, comme nous l'avons pu voir avec Perniola, il s'agit de termes à forte transparence sémantique, et porteurs aussi de facteurs culturels spécifiques. Néanmoins, ils ne sont pas présents dans les dictionnaires papier. Donc, il semble que l'importance de Popper, contrairement à celle de Freud, ait échappée aux lexicographes italiens, pour des raisons que nous trouvons difficiles à expliquer. Popper est toutefois en bonne compagnie, vu qu'un traitement similaire a été réservé à beaucoup d'autres personnalités du XXème siècle, comme par exemple Wittgenstein, Schopenhauer, Chomsky et Heidegger. La tendance à sous-estimer la pensée et la langue de ces (et d'autres) grands personnages a une forte portée culturelle, puisqu'elle semble tracer une ligne de démarcation nette entre les termes réels d'une langue, i.e. les mots comme ceux de la psychanalyse qui doivent ou peuvent être décrits dans les dictionnaires, et ceux virtuels, i.e. les mots du faillibilisme qui existent dans une langue mais qui ne sont ni lemmatisés, ni décrits. Les conséquences de tout cela sont plutôt importantes: si les termes utilisés par un auteur nous aident à comprendre sa pensée, le fait qu'ils ne soient pas lemmatisés ni donc divulgués par les dictionnaires empêche que l'auteur devienne aussi connu du grand public. Nous devons aussi noter que souvent par exemple un philosophe ne s'occupe pas seulement de philosophie ou un linguiste de linguistique. Parfois il peut écrire sur des arguments de grande actualité et d'intérêt général, comme Chomsky le fait aujourd'hui par rapport à la politique des Etats Unis ou comme Popper l'a fait peu avant sa disparition en écrivant sur le rapport entre société et télévision⁵⁴. Si l'on ne connaît pas un auteur et les questions qu'il traite, cela signifie aussi que l'on n'arrive pas à connaître des opinions importantes sur beaucoup d'arguments d'actualité. Cette forme d'ignorance induite a dans la *fiction* un antécédent plutôt sinistre et fameux. En fait, plus de cinquante ans auparavant, dans son roman *1984* l'écrivain anglais George Orwell avait illustré l'élaboration de techniques sophistiquées de contrôle sur l'intelligence humaine, en indiquant le strict rapport existant entre ces techniques, le langage et le lexique:

"L'objectif de la Néolangue était non seulement celui de fournir un moyen d'expression pour la conception du monde et pour les habitudes intellectuelles typiques

⁵⁴ Popper, K. R., Condry, J. 1996.

des adeptes du Socing, mais surtout de rendre impossible toute forme différente de pensés. Une fois que la Néolangue était définitivement adoptée et l'Archéolangue par contre définitivement oubliée, il était sous-entendu qu'une pensée hérétique (c'est-à-dire une pensée contrastant les principes du Socing) aurait été impensable, au moins pour ce que la pensée dépend des mots avec lesquels elle est en mesure d'être extériorisée. Son lexique était constitué de façon à permettre une expression exacte et assez subtile de chaque sens qu'un membre du Parti pouvait proprement désirer laisser entendre. Mais il excluait, en même temps, tous les autres sens possibles, et avec ceux-ci aussi la possibilité de les atteindre indirectement. Cela avait été obtenu pour une part grâce à la création de nouveaux mots, mais surtout grâce à la suppression de mots indésirables et à l'élimination de ces sens hétérodoxes qui avaient pu subsister et, pour ce qui était possible, de tous les sens en quelque mesure secondaires. Nous donnerons un seul exemple. Le mot libre existait toujours en Néolangue, mais il pouvait être utilisé seulement dans des phrases comme "Ce chien est libre de chiques" ou bien "Ce champ est libre de mauvaise herbe". Mais il ne pouvait pas être utilisé dans le vieux sens de "politiquement libre" ou "intellectuellement libre" vu que la liberté politique et intellectuelle n'existait guère, même pas comme concept, et que donc elle n'avait plus, par nécessité, un mot par lequel être désignée.”⁵⁵

Dans la peut-être utopique et sûrement distopique société orwellienne, l'application des principes du *Big Brother* avait porté à la création d'une nouvelle langue, précisément nommée *Néolangue*, du lexique de laquelle on avait éliminé tous les mots qui pouvait exprimer des idées potentiellement dangereuses pour l'ordre établi. En éliminant donc des mots spécifiques, on éliminait aussi les concepts qu'ils exprimaient; un tel aplatissement de la langue correspondait à un aplatissement

⁵⁵ Orwell, G. 1984 page 331 (c'est nous qui traduisons).

expressif égal par rapport à la conception du monde et à la liberté psychologique des êtres humains⁵⁶.

La situation que nous venons de décrire pour les dictionnaires papier n'a sûrement pas une telle portée idéologique, mais elle a des conséquences très similaires à celles indiquées par Orwell. En outre, d'un point de vue plus spécifiquement linguistique, les descriptions inexactes que l'on trouve dans les ouvrages papier mettent drastiquement à zéro l'opposition saussurienne entre *langue* et *parole*, et aussi la distinction chomskienne entre *compétence* et *performance*, précisément parce que la *parole* et la *performance*, respectivement partie individuelle et créative du langage, sont sacrifiées au nom d'une normalisation qui est linguistique au début, mais qui ensuite peut devenir psychologique et culturelle.

⁵⁶ Ce même passage est utilisé par Pinker (1994) comme une occasion pour démentir les thèses du déterminisme et relativisme linguistique nées à partir de l'hypothèse Sapir-Whorf. Pinker affirme que les êtres humains ne pensent pas en langage naturel, c'est-à-dire en italien, anglais ou français, mais qu'ils utilisent un langage de la pensée, qui Pinker appelle "langue de l'esprit" et qui sert à associer les images aux concepts avant de les exprimer verbalement. Pour Pinker, la pensée et ses concepts sont donc antécédents aux mots - on ne peut pas communiquer sans penser - et pour cette raison même si le terme justice était effacé de tous les lexiques du monde, l'idée et le concept de justice resteraient vivants dans la "langue de l'esprit", ce qui amènerait les humains à créer d'autres mots avec lequel les indiquer. En outre, la production de ces néologismes requerrait l'intervention d'une seule génération de locuteurs et serait confiée à la créativité linguistique des enfants. Certes, cette théorie de Pinker est amplement partageable, mais il est très probable qu'elle est valable seulement pour ces concepts qui comme justice, liberté, amour font depuis longtemps partie du patrimoine des êtres humains, naissent de leurs sentiments et de leur émotivité et, pour quelques, aspects ne dépendent pas seulement du niveau culturel et de la maturité des personnes individuellement. A notre avis, la théorie de Pinker ne serait toutefois pas réaliste pour des approches cognitives et épistémologiques tels que le faillibilisme et la psychanalyse, dont l'intelligence dépend beaucoup du niveau culturel des individus, et qui, en outre, n'incluent pas un seul concept mais une série de concepts souvent connectés l'un à l'autre. Ces mêmes concepts sont d'ailleurs le résultat d'un long et lent parcours fait par la recherche scientifique. Donc, pour un mot du faillibilisme ou de la psychanalyse éventuellement effacé, le processus de reconstruction serait sûrement plus complexe et tardif parce qu'il devrait parcourir de nouveau et obligatoirement tous les passages – historiques et aussi culturels – qui ont porté à sa première définition.

3.1.2 Quelle fin ont eu les verbes pronominaux?

Dans les dictionnaires, et plus généralement dans les grammaires et les œuvres sur la langue italienne, les agglutinations⁵⁷ du type verbe+clitique n'ont jamais été un prépondérant argument d'étude. En fait, sont assez limitées les pages dédiées à ces formes, et dans les dictionnaires papier beaucoup de formes pronominales ne sont pas signalées, même si elles sont normalement employées dans la langue soit orale qu'écrite, et que d'un point de vue syntaxique elles sont très régulières.

Néanmoins, nous avons de bonnes raisons pour nous intéresser à ce phénomène. Récemment, ces raisons ont été mises en évidence par l'analyse automatique des textes, qui permet le dépouillement rapide de corpus très vastes, grâce à l'apport d'applications en linguistique-informatique parfois plutôt complexes. Ces nouvelles procédures de récupération des informations ont permis, par exemple, d'établir qu'en italien, les occurrences des verbes pronominaux et des agglutinations peuvent être dénombrées avec précision et qu'elles sont typologiquement variées et suffisamment fréquentes en termes statistiques, ce qui ne permet pas de les laisser passer inaperçues⁵⁸.

Nous analyserons ces mots de l'italien, en étudiant leurs formes et usages sur la base des critères théoriques et pratiques du lexique-grammaire, i.e. une méthode de formalisation du langage naturel créée en France par Maurice Gross⁵⁹ et développée pour l'italien par Annibale Elia⁶⁰, Emilio D'Agostino, Maurizio Martinelli et Simona Vietri. En particulier, nous nous concentrerons sur les modalités de reconnaissance

⁵⁷ En réalité, les agglutinations sont un phénomène typique de l'italien mais qui existe aussi dans d'autres langues indo-européennes. Pour donner des exemples du faible intérêt suscité par ce phénomène, cfr. les entrées *enclisi* (enclise) et *enclitiche* (enclitique) de Serianni (1989), qui se limite à le signaler, sans en déterminer la syntaxe. Pour ces mêmes entrées et pour l'entrée *clitici* (clitiques), nous observons que des références limitées sont données dans AA. VV. (1973), peut-être à cause du caractère plurilingue de l'œuvre. Au contraire, Beccaria (1994) concentre la description de ce phénomène à l'intérieur de la seule entrée *enclisi*, dans laquelle il explique la valeur de termes tels que *clitico* (clitique), *gruppo clitico* (groupe clitique), *particella* (particule) et *proclisi* (proclise).

⁵⁸ En fait, dans des études plutôt récentes, nous avons pu vérifier que dans le corpus électronique d'un seul megabyte, composé principalement d'articles de journaux, il est possible de trouver 465 occurrences du type verbe+clitique.

⁵⁹ Voir Gross, M. 1975.

⁶⁰ Voir EMDA, 1981, et Elia, A. 1984.

automatique adoptées pour ces mots, en utilisant le logiciel Intex® et la théorie des automates à états finis sur laquelle il est basé.

En outre, nous verrons que souvent, dans les dictionnaires, il n'est pas possible de trouver des indications relatives à l'existence et à l'usage des verbes pronominaux, verbes qui sont la source syntaxique principale des agglutinations de l'italien.

3.1.1.1 Le système Intex® en italien

Nous avons montré quelques options spécifiques de ce système, telles que l'usage des automates à états finis pour la flexion automatique du DELAS en DELAF. Outre la gestion et l'application de moteurs linguistiques, Intex® a d'autres options de reconnaissance textuelle automatique qui se fondent sur l'usage des graphes et des automates à états finis (dorénavant indiqués par FSA). Comme nous le verrons, les FSA d'Intex® sont des instruments très ductiles et puissants, capables de reconnaître et d'étiqueter avec une précision extrême même les formes lexicales les plus particulières, et donc aussi les agglutinations.

Nous décrirons ce système dans l'optique de la linguistique théorique, et nous verrons comment quelques-unes de ces applications préliminaires peuvent nous aider soit à mettre à jour nos dictionnaires électroniques, soit à définir avec une précision majeure les bornes des dictionnaires papier.

D'un point de vue conceptuel, les dictionnaires électroniques, tels que ceux que nous sommes en train de décrire ici, et Intex®, qui est pour ces dictionnaires le contexte applicatif naturel, sont issus directement du lexique-grammaire, méthode pour laquelle l'unité linguistique minimale est la phrase simple⁶¹, i.e. un contexte grammatical formé par un seul élément prédicatif (un verbe, mais aussi ou un nom ou un adjectif) et par ses compléments essentiels. En général, l'étude des phrases simples d'une langue se fait surtout en analysant:

⁶¹ Pour la définition de la phrase simple voir Gross, M. 1980. Pour ce qui concerne les agglutinations et leur création, par rapport à la phrase simple, nous précisons que dans notre Nous nous occuperons exclusivement des agglutinations qui se vérifient dans et à partir des phrases simples de l'italien. Nous ne n'occuperons pas des agglutinations qui figurent dans et à partir des discours. Notre choix est motivé par le fait qu'une phrase simple, grâce à sa conformation syntaxique prédictible, permet de pronostiquer avec précision quelles formes agglutinées pourraient dériver d'elle. Les discours, au contraire, ne donnent pas les mêmes garanties, vu que leur formation est soumise à la créativité des locuteurs et donc est, en termes syntaxiques, moins prévisible.

- les règles de restriction de sélection et de distribution induites par les prédicats⁶²;
- les propriétés transformationnelles de chaque type de phrase simple.

Ces deux opérations permettent de repérer un nombre fini de classes syntaxiques prototypales, chacune d'entre elles contenant seulement des éléments prédicatifs à caractéristiques identiques.

Par exemple, si nous considérons le verbe italien *andare* (aller) et l'une de ses phrases simples, telle que:

Max va a Roma
(Max va à Rome)

il sera possible de formaliser la structure syntaxique suivante:

$N_0 V Loc N_I$

qui doit être lue comme suit: N_0 est le sujet du verbe, V le verbe, Loc indique la préposition locative et N_I le complément essentiel du verbe V . De même structure syntaxique, et donc de la même classe que *andare* sont des verbes comme *correre* (courir), *filare* (filer) et *volare* (voler) quand ils sont utilisés dans le même signifié que *andare*, comme dans les phrases suivantes:

⁶² Voir Gross, M. *ibid.*

Max (corre+scappa+vola) a Roma
(Max (court+fuit+vole) à Rome)

D'un point de vue transformationnel, nous pouvons observer que les phrases précédentes acceptent la pronominalisation du N_I :

Max (corre+scappa+vola) a Roma
(Max (court+fuit+vole) à Rome)

[T Pron]=:

Max (ci + vi) (corre+scappa+vola)
(Max y (court+fuit+vole))

Les classes structurelles ainsi repérées sont ensuite décrites à l'aide de tables syntaxiques, c'est-à-dire des matrices binaires dans lesquelles la présence et/ou l'absence de propriétés spécifiques distributionnelles et transformationnelles sont marquées avec les signes + ou -.

Le lexique-grammaire prévoit donc une description taxinomique des unités lexicales simples et des parties de grammaire qui leur sont associées. A part les déjà cités mots simples et composés, cette méthode a localisé d'autres types d'unités lexicales sémantiquement autonomes, c'est-à-dire les verbes composés et les expressions figées⁶³. A ces dernières typologies appartiennent des usages linguistiques tels que *andare bene* (aller bien), *andare male* (aller mal), *uscire dal seminato* (s'éloigner du sujet) et *arrampicarsi sugli specchi* (grimper sur les miroirs = défendre des cause perdues), i.e.

⁶³ Pour les verbs composés et les expressions figées de l'italien, voir Vietri, S. 1985, et Monteleone, M. 1989a et 1989b.

des expressions formées de plus d'un seul mot, dans leur ensemble avec une fixité structurale discrète ou forte et un sens global qui ne peut pas être obtenu en additionnant les sens des termes simples. Ces expressions, en outre, changent radicalement de sens, et peuvent même devenir non-grammaticales, si l'un des leurs éléments est remplacé

Par exemple, dans l'expression figée *arrampicarsi sugli specchi*, il n'est pas possible de modifier *specchi* ou substituer à ce complément *specchi* un autre mot, même sémantiquement proche, sans perdre le figement du signifié ou sans produire une phrase sémantiquement non acceptable:

?arrampicarsi sul vetro

(?grimper au verre)

?arrampicarsi sullo specchio

(?grimper au miroir)

Intex®, comme nous l'avons déjà indiqué, est un logiciel d'analyse textuelle automatique qui met en pratique les principes théoriques du lexique-grammaire et qui utilise les dictionnaires électroniques comme des moteurs linguistiques. Pour résumer ses fonctionnalités, nous dirons qu'à chaque fois que nous lui soumettons un texte à analyser, le programme compare les mots qui y sont avec ceux des dictionnaires électroniques. Ensuite, sur la base des indications morphologiques et grammaticales contenues dans les moteurs linguistiques, le programme crée des dictionnaires étiquetés du texte, séparés sur la base des unités lexicales reconnues. Ensuite, une option permet de mettre en évidence ces mots du texte qui ne sont pas présents dans les dictionnaires électroniques. Parmi ces mots, il sera possible de choisir ceux qui pourront être utilisés pour la mise à jour des même dictionnaires électroniques, qui, comme nous l'avons déjà vu, ne sont pas exhaustifs mais visent à l'exhaustivité.

La création des dictionnaires étiquetés des textes représente pour le logiciel le point de départ pour la gestion d'applications plus complexe, telles que celle des

automates et des transducteurs à états finis pour l'élaboration de concordances, d'analyses morphosyntaxiques et de parsing. Pour ces opérations aussi, Intex utilise les indications morphologiques et grammaticales insérées dans les différents dictionnaires électroniques, qui, à part les indications flexionnelles comprennent pour les verbes: les étiquettes relatives à la transitivité, l'intransitivité et les auxiliaires, et pour les substantifs: des étiquettes sémantiques comme humain, non-humain et animé.

Donc, Intex® ne pourrait matériellement pas opérer sans les dictionnaires électroniques et pour cette raison ces derniers, avec tout l'ensemble du programme, sont un instrument des plus importants dans l'application, la vérification et la révision des théories linguistiques élaborées par le lexique-grammaire. Nous avons vu que les dictionnaires électroniques sont l'endroit où le lexique est étiqueté et commenté sur la base des usages réels de ses éléments. En utilisant Intex®, il est effectivement possible d'utiliser ces commentaires pour dépouiller les corpus et vérifier comment sont vraiment utilisés les prédicats d'une langue ou analyser aussi leurs occurrences textuelles à l'aide des concordances ou bien utiliser le logiciel pour localiser seulement dans un texte des constructions syntaxiques spécifiques, de façon à effectuer des analyses détaillées sur les propriétés de mots simples ou en séquence.

Grâce à leur structure rigoureusement formelle, même s'ils sont une création directe du lexique-grammaire, les dictionnaires électroniques ont été utilisés avec succès aussi dans d'autres domaines comme la génération de textes et la traduction assistée par l'ordinateur⁶⁴. Cette tendance en a confirmé la validité, surtout en termes d'amplitude des listes de lemmes, qui en ce qui concerne la langue italienne sont fort probablement les plus complets. Néanmoins, nous ne devons plus oublier que les dictionnaires électroniques visent à l'exhaustivité mais ne sont pas exhaustifs parce qu'ils ne contiennent pas tous les mots potentiellement utilisables dans une langue.

Un cas particulier de mots qui ne sont pas répertoriés dans nos dictionnaires électroniques sont les agglutinations, que nous allons analyser en détail. Dans le DELAS, nous observons que les seules agglutinations listées sont celles relatives aux verbes pronominaux à l'infinitif, telles que:

⁶⁴ Voir AA. VV. 1989.

aggrapparsi (s'agripper)
arrampicarsi (grimper)
genuflettersi (s'agenouiller)
inginocchiarsi (s'agenouiller)

qui, en italien, ne sont pas des lemmes d'usage autonome, mais figurent seulement dans quelques constructions causatives⁶⁵.

D'autres types de pronominaux, tels que:

baciarsi (s'embrasser)
redimersi (se délivrer)
pettinarsi (se coiffer)

sont en rapport de dérivation morphologique avec les formes respectives non-pronominales ou canoniques, à partir desquelles ils sont formés en ajoutant à la fin du mot la particule pronominale *si* (se) et en élidant la dernière lettre de l'infinitif. Si nous observons les éléments de la liste précédente, nous dirons que *baciarsi* dérivera de *baciare* (embrasser), *redimersi* de *redimere* (délivrer) et *pettinarsi* de *pettinare* (coiffer). Dans le DELAS, il sera suffisant de lister seulement ces formes non-pronominales, et de faire en sorte qu'à partir de celles-ci soient créées toutes les formes pronominales correspondantes, en utilisant les propriétés syntaxiques des verbes de départ et en créant des transducteurs à états finis spécifiques à l'aide de Intex®.

D'ailleurs, la méthode de flexion automatique des verbes pronominaux est différente de celle déjà illustrée par rapport aux mots, simples et composés, parce qu'elle est structurée et appliquée sur deux phases différentes, sur la base des formes des mêmes verbes, qui formellement peuvent figurer:

⁶⁵ Par exemple, dans la phrase *Max fa aggrappare Luca al suo braccio* (Max fait agripper Luc à son bras). Pour cette raison, en italien ces formes sont appelées infinitives causatives.

- comme mot unique, tels que par exemple *aggrappati* (agrippe-toi), *arràmpicati* (grimpe-toi), *genuflettiti* (agenouille-toi), *inginocchiati* (agenouille-toi), *redimiti* (délivre-toi);
- comme deux mots séparés, lorsque, pour des raisons syntaxiques, il est nécessaire de disjoindre les clitiques de la forme verbale, comme par exemple avec *mi aggrappo* (je m'agrippe), *si arrampica* (il grimpe), *ti genufletti* (tu t'agenouilles), *ci inginocchiamo* (nous nous agenouillons), *si redimono* (ils se délivrent).

Comme il s'agit de formes simples, les mots uniques sont à juste titre prévus par la flexion du DELAS et donc insérés dans le DELAF. Outre l'infinitif et l'infinitif causatif, la typologie de ces mots comprend:

- les deuxièmes personnes du singulier et du pluriel de l'impératif présent, comme par exemple les déjà cités *aggràppati*, *arràmpicati*, *genuflettiti*, *inginocchiati*;
- toutes les entrées du participe présent, comme *aggrappatemi* (m'agrippant) *inginocchiantici* (nous agenouillant);
- toutes les entrées du participe passé, comme *aggrappatasi* (s'étant agrippée), *inginocchiatomi* (m'étant agenouillé);
- toutes les entrées du gérondif, comme *aggrappandosi* (en s'agrippant), *inginocchiandomi* (en m'agenouillant).

Au contraire, les formes verbales pronominales non uniques sont insérées dans le DELAFC, vu qu'il s'agit de formes verbales composées comme le démontrent aussi les phrases simples suivantes:

- 1) *Max e Lucia si baciano*
(Max et Lucia s'embrassent)
- 2) *Io mi redimo*
(Je me délivre)
- 3) *Tu ti pettini (E+_i capelli)*
(Tu te coiffes (E+les cheveux))

Il est donc possible d'effectuer une distinction ultérieure par rapport à ces formes en ce qui concerne la différence existant entre agglutinations morphologiques et agglutinations syntaxiques. Les premières sont essentiellement des entrées fléchies d'un verbe pronominal qui ne peuvent pas être obtenues différemment sauf en agglutinant la particule au verbe. Cela arrive par exemple avec des formes comme:

inginocchiandomi (en m'agenouillant)

qui ne peut être écrit comme suit:

**inginocchiando me stesso* (*en agenouillant moi-même)

Il existe d'autres formes agglutinées qui ne sont pas morphologiquement prévisibles, parce qu'elles ne peuvent être obtenues par flexion automatique et donc ne sont pas insérées dans le DELAF et le DELACF. Nous faisons référence à des mots tels que:

abbracciarlo (l'embrasser)
capirne (en comprendre)
inviandocelo (en nous l'envoyant)

qui respectivement ne dérivent pas des mots suivants:

abbracciarsi (s'embrasser)
capirsi (se comprendre)
inviarsi (s'envoyer)

mais de pronominalisations appliquées à des phrases simples du type:

Max vuole abbracciare Luca
(Max veut embrasser Luca)

[T Pron]=:

Max vuole abbracciarlo
(Max veut l'embrasser)

dans laquelle le verbe *abbracciare* figure dans sa forme transitive et non sous forme pronominale. Encore, nous pouvons avoir:

Eva non riesce a capire una parte del discorso

(Eva n'arrive pas à comprendre une partie du discours)

[T Pron]=:

Eva non riesce a comprenderne una parte

(Eva n'arrive pas à en comprendre une partie)

et aussi:

Luca e Paolo stanno inviando il documento a noi

(Luca et Paolo sont en train de nous envoyer le document)

[T Pron]=:

Luca e Paolo stanno inviandocelo

(Luca et Paolo sont en train de nous l'envoyer)

Les agglutinations syntaxiques, que nous distinguerons donc des agglutinations morphologiques, sont créées presque toujours à partir des phrases simples, et s'obtiennent grâce à la pronominalisation des compléments directs ou indirects des verbes qui figurent dans les mêmes phrases, et sont souvent des transformations libres, non imposées par des restrictions grammaticales ou syntaxiques: pour ces raisons elles sont presque toujours liées à l'emphase ou au style des locuteurs. Nous vérifierons comment il est possible de localiser automatiquement ces formes dans des textes, grâce à une étude syntaxique préliminaire des verbes qui y participent, et nous verrons

qu'Intex® et les dictionnaires électroniques nous donneront un apport considérable. Néanmoins, l'existence de ces agglutinations syntaxiques nous porte à affirmer que l'étude de toutes les formes agglutinées italiennes devra nécessairement être effectuée sur trois niveaux différents, en l'occurrence:

- un niveau morphologique, utile pour individualiser ces formes d'un verbe pronominal donné qui peuvent être obtenues à l'aide de la flexion automatique;
- un niveau syntaxique, dans lequel l'intervention humaine est prépondérante pour prévoir et reconnaître seulement les agglutinations syntaxiquement grammaticales;
- un niveau formel, i.e. un niveau de reconnaissance pratique et automatique des agglutinations, dont se chargent Intex® et les FSA et qui se base: soit sur les résultats du niveau morphologique, soit sur les résultats de l'analyse syntaxique des verbes.

Pour démontrer combien il est important de mener à terme et de façon détaillée toutes ces approches, essayons d'imaginer que, pendant la lecture automatique d'un texte à l'aide D'Intex® nous demandons au logiciel d'effectuer l'opération suivante:

*Etiqueter chaque mot qui termine avec **rne** comme un verbe à l'infinitif agglutiné au clitique **ne**.*

Par cette opération, nous obtiendrons que des mots tels que:

abbracciarne (en embrasser)

ricordarne (en rappeler)

salvarne (en sauver)

seront correctement étiquetés comme *verbe à l'infinitif agglutiné au clitique **ne***, mais la même étiquette sera attribuée de façon erronée à d'autres mots, c'est-à-dire des substantifs, comme par exemple:

carne (viande)

tagliacarne (coupe-viande)

tritacarne (hache-viande)

Une autre instruction comme celle-ci:

Etiqueter chaque mot qui termine en **ine** comme un verbe à l'impératif, deuxième personne singulière, agglutiné au clitique **ne**.

étiquetterait correctement les agglutinations suivantes:

prèndine (prends en)

spìngine (pousse-z-en)

vivine (vis en)

mais erronément des substantifs tels que:

regine (reines)

spine (épines)

vetrine (vitrines)

Les instructions précédentes, exclusivement formelles, contraignent donc le logiciel à accomplir des opérations non valables voire inutiles, et elles démontrent que l'approche formelle est fautive si on ne lui associe pas les informations morphologiques, lexicales et syntaxiques adéquates à la reconnaissance correcte des séquences. Comme nous le verrons, c'est seulement en affinant les instructions de reconnaissance qu'il sera possible de réduire sensiblement la marge d'erreur. En ce sens, un apport important nous sera offert par le lexique-grammaire qui est l'une des rares méthodes linguistiques en mesure de définir avec précision les propriétés distributionnelles des verbes et aussi, pour les agglutinations, toutes les pronominalisations grammaticales possibles des phrases simples.

Après avoir traité la partie morphologique de la description, il faudra effectuer une étude grammaticale et syntaxique des agglutinations comme formes transformées d'un verbe.

3.1.1.2 Les agglutinations de l'italien

Il est bien connu qu'en italien une agglutination représente un mot simple composé d'une forme verbale et d'une ou de plusieurs particules pronominales, dites aussi clitiques. La liste complète de ces particules est résumées dans la table ci-dessous:

ce	ci	-	-
gli	glie	-	-
la	le	li	lo
me	mi	-	-
ne	-	-	-
se	si	-	-
te	ti	-	-
ve	vi	-	-

Table 7

Exception faite pour quelques entrées des verbes pronominaux, les agglutinations sont des transformations syntaxiquement non nécessaires, presque toujours liées à l'emphase ou au style des locuteurs. Elles peuvent être obtenues en pronominalisant les compléments directs ou indirects des verbes. Si nous considérons la phrase simple:

1. *Bruciate la casa*
(Brûlez la maison)

et en appliquant la pronominalisation du complément indirect *la casa*, nous obtenons l'agglutination:

1a. [T Pron] *Bruciatela*
(Brûlez-la)

qui est composé de *bruciate*, première personne pluriel de l'impératif de *bruciare* (brûler) et du clitique *la* (la), sur la base de la correspondance:

la = *la casa*
(la = la maison)

Comme nous l'avons vu, par rapport aux agglutinations morphologiques, celles qui sont syntaxiques ne concernent pas toutes les entrées ou tous les modes verbaux, mais seulement quelques-uns d'entre eux, c'est-à-dire:

- l'infinitif, comme dans *bruciar mi* (me brûler), *bruciar li* (les brûler)
- la deuxième personne du singulier et du pluriel de l'impératif, comme dans *brucialo* (brûle-le) et *bruciateli* (brûlez-les);

- le participe présent, comme dans *bruciatesi* (qui se sont brûlés), *bruciantisi* (qui sont en train de se brûler);
- le participe passé, comme dans *bruciatosi* (qui s'est brûlé), *bruciatasi* (qui se sont brûlés);
- le gérondif, comme dans *bruciandolo* (en le brûlant), *bruciandola* (en la brûlant).

En outre, il n'est pas possible d'agglutiner aux formes verbales plus de deux clitiques, et ceci indépendamment du nombre de compléments qui sont à pronominaliser. Par exemple, si nous considérons le verbe transitif *dare* (donner) dans la phrase:

2. *Max dà un pezzo di crosta di pane a me*
(Max me donne un morceau de croûte de pain)

qui peut être classé comme verbe transitif avec une structure du type:

N0 V N1 PREP N2

et si nous appliquons à la phrase (2) toutes les pronominalisations possibles relative à N₁ et à N₂, bien que l'N₁ ait une structure syntagmatique, avec le gérondif nous pouvons obtenir seulement deux agglutinations grammaticales:

[T Pron] 1a. *Dandomelo* (en me le donnant)

[T Pron] 1b. *Dandomene un pezzo* (en me donnant un morceau)

dont la deuxième aussi est paraphrastique. L'ajout d'un troisième clitique aux deux déjà présents dans (1a) et (1b) ne produit pas de formes grammaticales:

[T Pron] 1c. **Dandomelone*

[T Pron] 1d. **Dandomenelo*

D'un point de vue morphologique, dans les agglutinations syntaxiques aussi nous observons l'élision du *e* final des verbes à l'infinitif qui sont associés à un clitique, comme il est possible de noter dans les exemples suivants:

2. *calpestare+le* = *calpestarle* (les piétiner);

3. *correggere+li* = *correggerli* (les corriger);

4. *esprimere+si* = *esprimersi* (s'exprimer).

Toujours d'un point de vue morphologique, il faut souligner que quelques entrées monosyllabiques des verbes *andare* (aller), *dare* (donner), *dire* (dire), *fare* (faire) et *stare* (rester), i.e. ceux de la deuxième personne du singulier de l'impératif, prennent part aux agglutinations en redoublant la consonne initiale des clitiques *ci*, *mi*, *ti*, *lo*, *la*, *li*, *le*, *ne*. Ces formes perdent aussi la dernière lettre – ou l'apostrophe, si présent – et l'accent final :

5. *(dai+dài+dà+ da')*+*mi* = *dammi* (donne-moi)

6. *(fai+fà+fa')*+*le* = *falle* (fais-lui)

7. *(vai+vài+và+va')*+*lo* = *vallo* (vas-le)

8. *(dì+di')*+*ci* = *dicci* (dis-nous)

9. *(stai+sta+sta')*+*ne* = *stanne* (reste-z-en)

Un redoublement identique survient quand les mêmes entrées sont associées aux clitiques *ce*, *me*, *te* suivis d'un second clitique:

10. $(dai+d\grave{a}i+d\grave{a}+da') + me + ne = damm\grave{e}ne$

11. $(fai+f\grave{a}+fa') + ce + lo = fa\grave{c}celo$

12. $(vai+v\grave{a}i+v\grave{a}+va') + te + la = vattela$

13. $(d\grave{i}+d\grave{i}') + ce + ne = dicc\grave{e}ne$

14. $(stai+sta+sta') + te + ne = stattene$

Finalement, dans la formation de quelques agglutinations, nous observons l'application de l'accord en genre et en nombre entre verbes et compléments pronominalisés, i.e. entre verbes et particules à l'intérieur des mêmes agglutinations. Par exemple, dans les deux phrases suivantes:

15. Max ha bruciato il tavolo (Max a brûlé la table)

16. Max ha bruciato le case (Max a brûlé les maisons)

les participes passés des deux verbes sont à l'origine des agglutinations suivantes:

15a. *bruciatolo* (après l'avoir brûlé)

16a. *bruciatele* (après les avoir brûlées)

dans lesquelles soit les entrées verbales, et les particules, en termes de genre et de nombre, créent un processus déictique et extra-contextuel vers un complément masculin, dans le premier cas, et féminin pluriel, dans le deuxième. La mise à zéro de

l'accord dans les formes agglutinées rendrait impossible le parcours à rebours vers les phrases simples qui les ont engendrées.

Il est bien connu que le participe passé est une entrée verbale qui fléchit ses formes en genre et aussi en nombre, et pour cette raison il est possible, sinon nécessaire de respecter la coréférence d'un verbe donné avec le genre et le nombre des compléments pronominalisés et présents dans les agglutinations. Donc, les agglutinations suivantes ne seront pas grammaticales:

**bruciatila* (après les avoir brûlée)

**bruciatilo* (après les avoir brûlé)

**bruciatalo*⁶⁶ (après la avoir brûlé)

⁶⁶ Nous observons qu'à partir de la phrase:

Max ha bruciato i pannelli (Max a brûlé les panneaux)

nous obtenons l'agglutination:

bruciatili (après les avoir brûlés)

tandis que n'est pas grammaticale l'agglutination:

**bruciatoli* (après les avoir brûlé).

Cela porterait à penser que le contrôle des relations de co-occurrence entre participe passé et clitique sont effectuées exactement pendant la phase de formation des agglutinations ou pendant les pronominalisations. Observons donc les phrases suivantes:

Max ha visto Lucia ed è partito
(Max a vu Lucie et il est parti)

qui transformée en phrase scindée devient:

(*E+Una volta*) *vista Lucia, Max è partito*
((E + Une fois) vue Lucie, Max est parti)

Cela ferait penser que la vérification des règles de co-occurrence entre verbe et compléments pronominalisés est aussi effectuée pendant les transformations passives. Néanmoins, nous avons:

Lucia è andata a casa
(Lucie est allée à la maison)

et non:

**Lucia è andato a casa*
(*Lucie est allé à la maison)

vue aussi la non grammaticalité de phrases comme:

17. **Max ha bruciata il pane*
 (*Max a brûlée le pain)

ce qui met en évidence un comportement différent des participes passés en présence d'un autre auxiliaire.

3.1.1.3 Automates à états finis et reconnaissance automatique des textes

Dans Intex®, un automate à états finis est une grammaire formelle qui peut être utilisée pour achever différents types de tâches, parmi lesquels il y a aussi la reconnaissance automatique des textes. D'un point de vue graphique, un automate se présente comme indiqué dans la figure suivante:

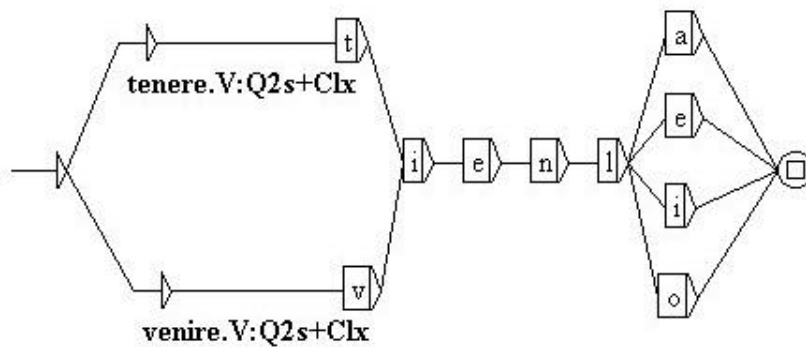


Figure 1

Il s'agit donc d'un graphe qui indique au programme le parcours de lecture, avec un état initial, un état final et une série de nœuds intermédiaires. Plus particulièrement, à partir du graphe de la figure 1 on crée l'automate à états finis qui reconnaît des agglutinations spécifiques, telles que:

vienlo (viens-le), *vienla* (viens-la), *vienli* (viens-les), *vienle* (viens-les)

tienlo (tiens-le), *tienla* (tiens-la), *tienli* (tiens-les), *tienle* (tiens-les)

et les étiquettes comme deuxièmes personnes du singulier de l'impératif, respectivement des verbes *venire* (venir) et *tenere* (tenir). Une fois reconnues, les séquences sont listées par Intex® dans les dictionnaires du texte analysé, i.e. celui qui a la même structure que le DELAF:

tienla,tenere.V:Q2s+Clx

tienle,tenere.V:Q2s+Clx

tienli,tenere.V:Q2s+Clx

tienlo,tenere.V:Q2s+Clx

vienla,venire.V:Q2s+Clx

vienle,venire.V:Q2s+Clx

vienli,venire.V:Q2s+Clx

vienlo,venire.V:Q2s+Clx

Dans l'exemple précédent, l'étiquette *Clx* indique la présence d'un clitique dans la voix verbale. L'automate de la figure (1), en format texte et donc de grammaire formelle, a la structure suivante:

```

11 13
%a%e%i%l%n%o%t%v%<E>/venire.V:Q2s+Clx%<E>%<E>/tenere.V:Q2s+Clx%
: 8 2 10 3 -1
: 7 4 -1
: 6 5 -1
: 2 6 -1
: 2 6 -1
: 1 7 -1
: 4 8 -1
: 3 9 -1
: 0 10 1 11 2 12 5 13 -1
t -1
t -1
t -1
t -1

```

Cette version de l'FSA n'est pas créée manuellement par l'utilisateur mais automatiquement, à l'aide d'une option spécifique d'Intex® qui convertit le graphe visible dans la figure (1), et a été réalisé par l'éditeur graphique du logiciel. Donc, avec Intex® il n'est pas nécessaire que l'utilisateur connaisse les règles de création des grammaires formelles pour pouvoir construire les FSA, mais il est suffisant qu'il sache utiliser l'éditeur graphique.

L'analyse des propriétés distributionnelles d'un verbe est nécessaire pour décider quelles sont les agglutinations que l'on doit faire reconnaître et étiqueter par un automate. Pour mieux exemplifier ces aspects, nous nous concentrerons ici sur les propriétés du verbe *bruciare* (brûler) et nous en analyserons les usages suivants:

- a. intransitif, comme dans la phrase *La casa di Paolo brucia* (La maison de Paul brûle);
- b. réciproque, comme dans la phrase *Max e Eva si bruciano* (Max et Eve se brûlent);

- c. réflexif de possession⁶⁷, comme dans la phrase *Max si brucia un dito* (Max se brûle un doigt);
- d. réflexif, comme dans la phrase *Max si brucia* (Max se brûle);
- e. transitif, comme dans la phrase *Max brucia la casa di Paolo* (Max brûle la maison de Paul).

Nous n'indiquerons pas en détail les contextes phrastiques dont dérivent les pronominalisations que nous observerons, mais nous focaliserons principalement notre intérêt sur la méthode de reconnaissance automatique et sur son efficacité⁶⁸. En réalité, la perception du signifié d'une agglutination est intuitive et seulement dans une deuxième phase elle est ramenée à la transformation qui l'a créée. Parmi les formes agglutinées que nous classerons, il y en a quelques-unes qui sont facilement acceptables, d'autres ambiguës, d'autres grammaticales mais utilisées très rarement.

Nous soulignons toutefois la particularité de la phrase (a) de la liste précédente, dont il n'est pas possible d'obtenir des agglutinations, puisque ce type d'usage intransitif ne contient pas de compléments indirects. Donc, la phrase (a) ne sera pas exemplifiée dans les graphes qui suivent.

⁶⁷ Le réflexif de possession est en réalité un usage transitif particulier, avec une structure du type:

N0 V N1

dans laquelle est vraie l'équation:

N1 := Npc de N0

où **Npc** signifie nom de partie du corps. En outre, nous soulignons que la forme de l'infinitif d'un verbe à cet usage est pronominal, comme avec *bruciarsi* (se brûler), tandis que la forme originnaire du complément direct est:

PossN0 Npc

comme le démontre l'équivalence:

Max si brucia un dito =: *Max brucia un suo dito* (Max se brûle un doigt =: Max brûle un de ses doigts).

⁶⁸ A ce propos, nous soulignons que les automates que nous allons montrer ont été appliqués avec succès et plusieurs fois à beaucoup de textes italiens.

a. *Bruciare* réciproque

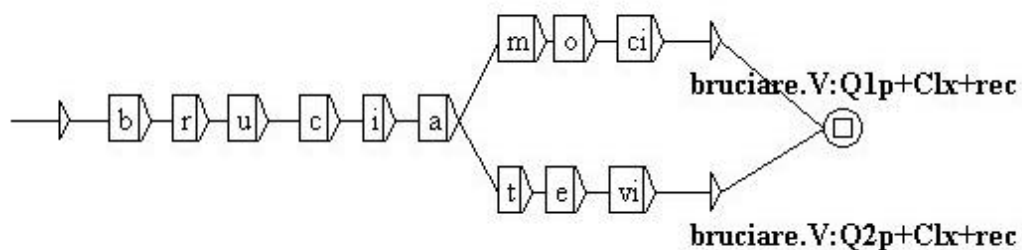


Figure 2: graphe pour l'impératif.

Ce graphe reconnaît et étiquette les agglutinations suivantes (2 formes):

bruciamoci, bruciatevi

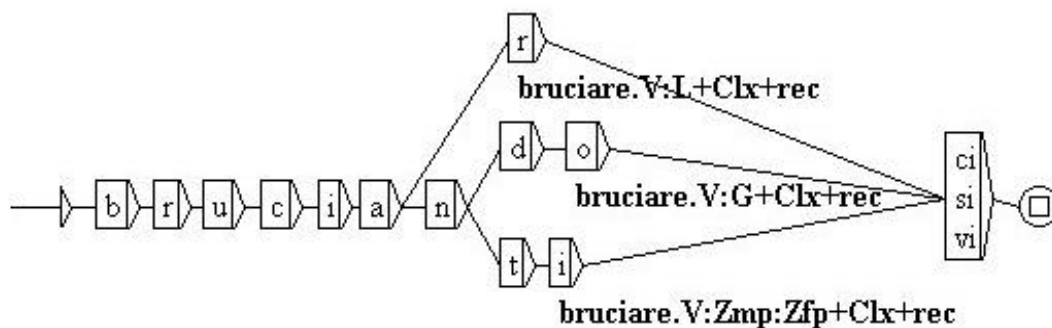


Figure 3: graphe pour l'infinitif, le gérondif et le participe présent.

Ce graphe reconnaît et étiquette les agglutinations suivantes (9 formes):

bruciarci, bruciarvi, bruciarci, bruciandoci, bruciandovi, bruciandosi, bruciantici, bruciantisi, bruciantivi

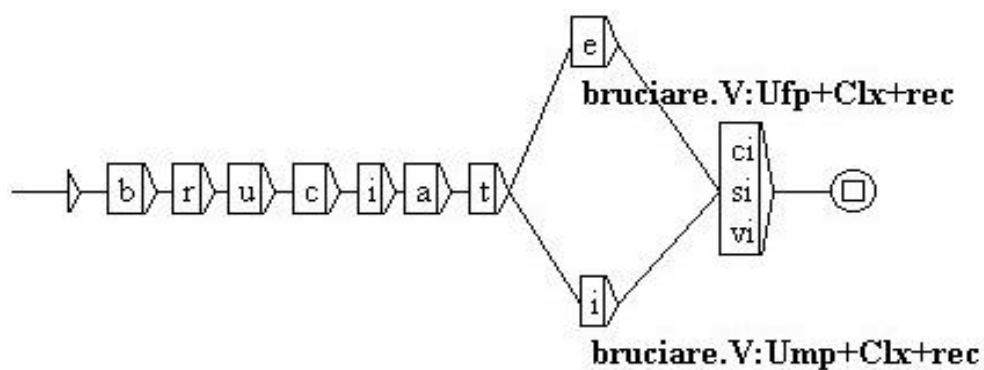


Figure 4: graphe pour le participe passé.

Ce graphe reconnaît et étiquette les agglutinations suivantes (6 formes):

bruciateci, bruciatesi, bruciatevi, bruciatoci, bruciatosi, bruciativi

b. *Bruciare* réflexif de possession

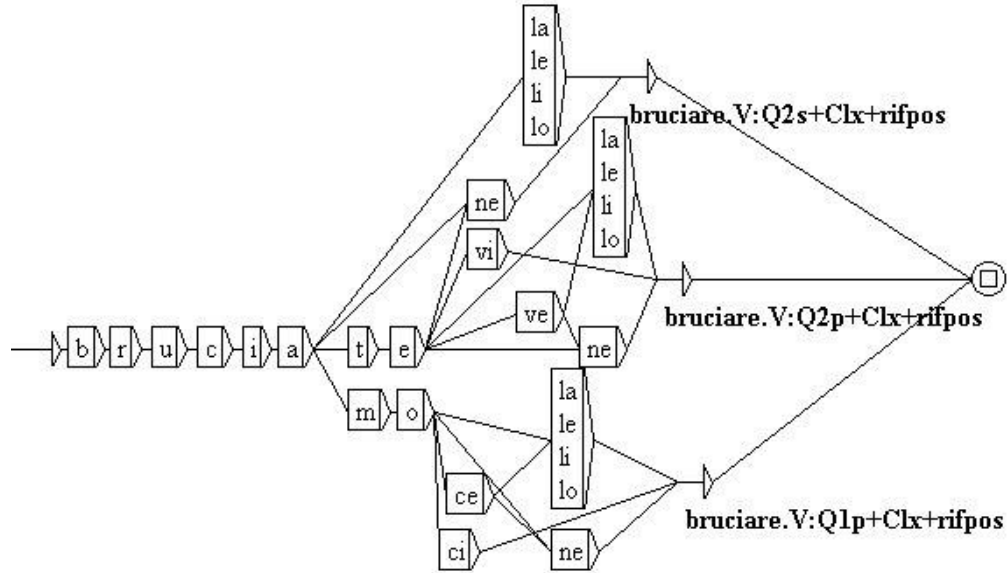


Figure 5: graphe pour l'impératif.

Ce graphe reconnaît et étiquette les agglutinations suivantes (32 formes):

bruciala, bruciale, bruciali, brucialo, bruciane, brùciatela⁶⁹, brùciatele, brùciateli, brùciatelo, brùciatene, bruciàtela, bruciàtele, bruciàteli, bruciàtelo, bruciàtene, bruciatevela, bruciatevele, bruciateveli, bruciatevelo, bruciatevene, bruciamola, bruciamole, bruciamoli, bruciamolo, bruciamone, bruciamocela, bruciamocele, bruciamoceli, bruciamocelo, bruciamocene, bruciamoci, bruciatevi

⁶⁹ L'accent a été inséré seulement dans le but de différencier la prononciation des homographes *brùciatela* (brûle-les toi) et *bruciàtele* (brûlez-les), qui sont deux agglutinations différentes.

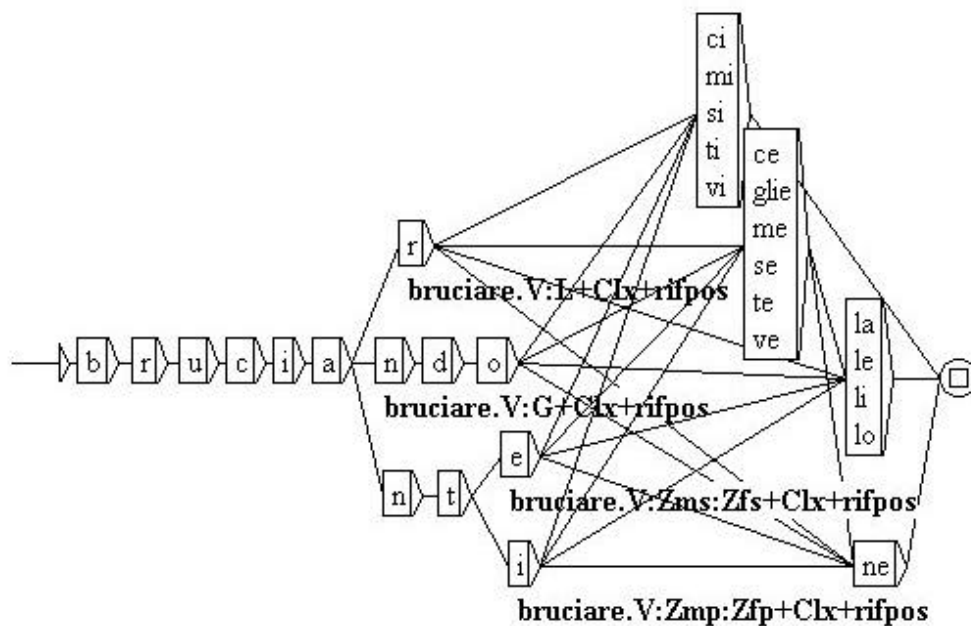


Figure 6: graphe pour l'infinitif, le gérondif et le participe présent.

Ce graphe reconnaît et étiquette les agglutinations suivantes (160 formes):

bruciarci, bruciarmi, bruciarvi, bruciarti, bruciarsi, bruciarselo, bruciarsela, bruciarseli, bruciarsele, bruciarsene, bruciandoci, bruciandomi, bruciandovi, bruciandoti, bruciandosi, bruciandoselo, bruciandosela, bruciandoseli, bruciandosele, bruciandosene, brucianteci, bruciantemi, bruciantevi, brucianteti, bruciantesi, brucianteselo, bruciantesela, brucianteseli, bruciantesele, bruciantesene, bruciantici, bruciantimi, bruciantivi, bruciantiti, bruciantisi, bruciantiselo, bruciantisela, bruciantiseli, bruciantisele, bruciantisene, bruciarlo, bruciarla, bruciarli, bruciarle, bruciarne, bruciarcelo, bruciarcela, bruciarceli, bruciarcele, bruciarcene, bruciarcelo, bruciarcela, bruciarceli, bruciarcele, bruciarcene, bruciarmelo, bruciarmela, bruciarmeli, bruciar mele, bruciar mene, bruciarglielo, bruciargliela, bruciarglieli, bruciargli ele, bruciargli ene, bruciar telo, bruciar tela, bruciar teli, bruciar tele, bruciar tene, bruciarvelo, bruciarvela, bruciarveli, bruciarvele, bruciarvene,

bruciandolo, bruciandola, bruciandoli, bruciandole, bruciandone,
bruciandocelo, bruciandocela, bruciandoceli, bruciandocele,
bruciandocene, bruciandomelo, bruciandomela, bruciandomeli,
bruciandomele, bruciandomene, bruciandoglielo, bruciandogliela,
bruciandoglieli, bruciandogliele, bruciandogliene, bruciandotelo,
bruciandotela, bruciandoteli, bruciandotele, bruciandotene,
bruciandovelo, bruciandovela, bruciandoveli, bruciandovele,
bruciandovene, bruciantelo, bruciantela, brucianteli, bruciantele,
bruciantene, bruciantecelo, bruciantecela, brucianteceli, bruciantecele,
bruciantecene, bruciantemelo, bruciantemela, bruciantemeli,
bruciantemele, bruciantemene, brucianteglielo, bruciantegliela,
brucianteglieli, bruciantegliele, bruciantegliene, bruciantetelo,
bruciantetela, brucianteteli, bruciantetele, bruciantetene,
bruciantevelo, bruciantevela, brucianteveli, bruciantevele,
bruciantevene, bruciantilo, bruciantila, bruciantili, bruciantile,
bruciantine, brucianticelo, brucianticela, brucianticeli, brucianticele,
brucianticene, bruciantimelo, bruciantimela, bruciantimeli,
bruciantimele, bruciantimene, bruciantiglielo, bruciantigliela,
bruciantiglieli, bruciantigliele, bruciantigliene, bruciantitelo,
bruciantitela, bruciantiteli, bruciantitele, bruciantitene, bruciantivelo,
bruciantivela, bruciantiveli, bruciantivele, bruciativene

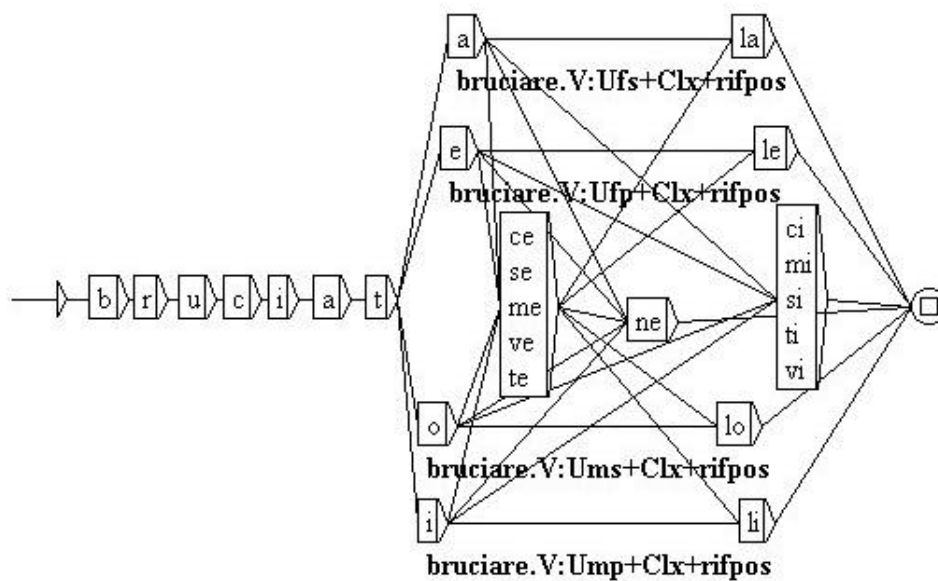


Figure 7: graphe pour le participe passé.

Ce graphe reconnaît et étiquette les agglutinations suivantes (68 formes):

bruciatami, bruciatati, bruciatasi, bruciataci, bruciatavi, bruciatomi, bruciatoti, bruciatosi, bruciatoci, bruciatovi, bruciatimi, bruciatiti, bruciatisi, bruciatichi, bruciativi, bruciatemi, bruciateti, bruciatesi, bruciateci, bruciatevi, bruciatata, bruciatele, bruciatili, bruciatolo, bruciatone, bruciatine, bruciatene, bruciatane, bruciatacela, bruciatasela, bruciatamela, bruciatavela, bruciatatela, bruciatacene, bruciatasene, bruciatamene, bruciatavene, bruciatatene, bruciatecele, bruciatesele, bruciatemele, bruciatevele, bruciatetele, bruciatecene, bruciatesene, bruciatemene, bruciatevene, bruciatetene, bruciatriceli, bruciatisel, bruciatimeli, bruciativeli, bruciatiteli, bruciatricene, bruciatisen, bruciatimene, bruciativene, bruciatitene, bruciatocelo, bruciatoselo, bruciatomelo, bruciatovelo, bruciatotelo, bruciatocene, bruciatosene, bruciatomene, bruciatovene, bruciatotene

c. *Bruciare* réflexif

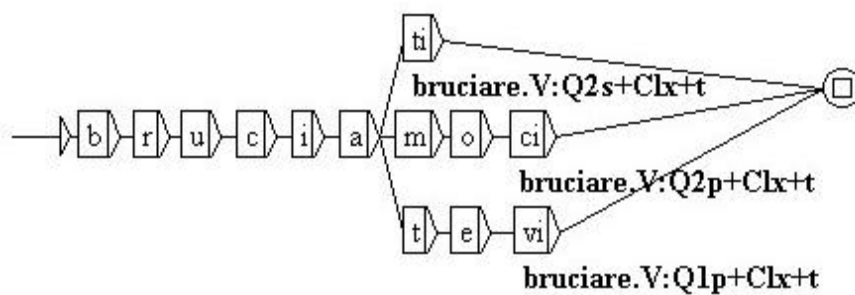


Figure 8: graphe pour l'impératif.

Ce graphe reconnaît et étiquette les agglutinations suivantes (3 formes):

*brùciati*⁷⁰, *bruciamoci*, *bruciatevi*

⁷⁰ Il faut noter l'homographie avec le participe passé *bruciàti* (brûlés).

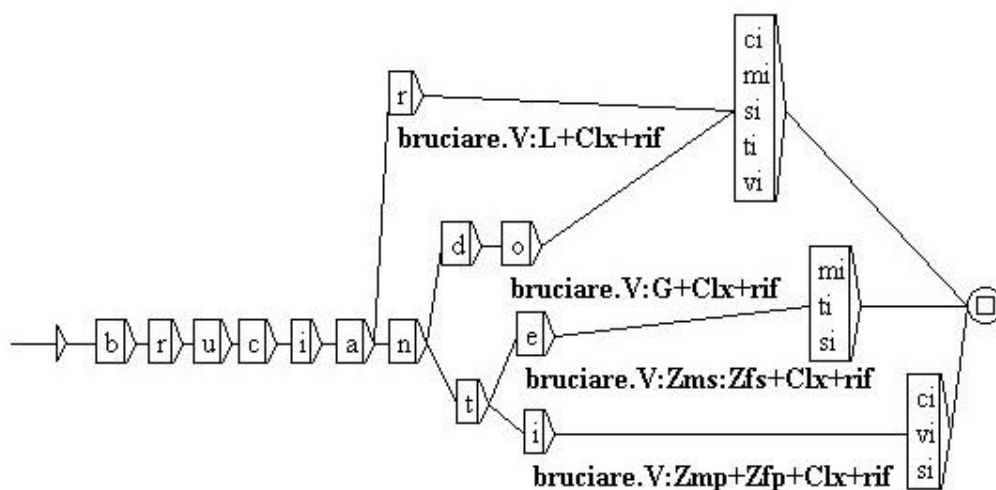


Figure 9: graphe pour l'infinitif, le gérondif et le participe présent.

Ce graphe reconnaît et étiquette les agglutinations suivantes (16 formes):

bruciarci, bruciarmi, bruciarci, bruciarti, bruciarvi, bruciandoci, bruciandomi, bruciandosi, bruciandoti, bruciandovi, bruciantemi, brucianteti, bruciantesi, bruciantici, bruciantivi, bruciantisi

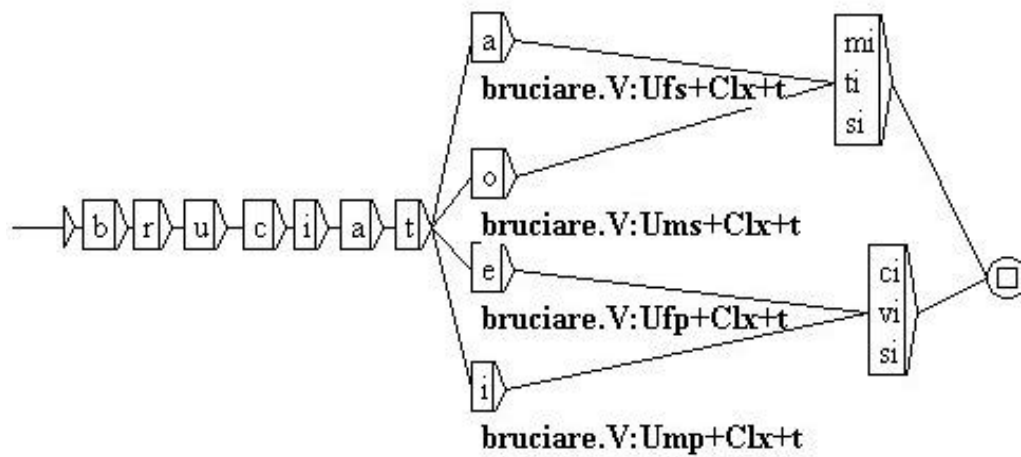


Figure 10: graphe pour le participe passé.

Ce graphe reconnaît et étiquette les agglutinations suivantes (12 formes):

bruciatami, bruciatati, bruciatasi, bruciatomi, bruciatoti, bruciatosi,
bruciateci, bruciatevi, bruciatesi, bruciatichi, bruciativi, bruciatisi

d. *Bruciare* transitif

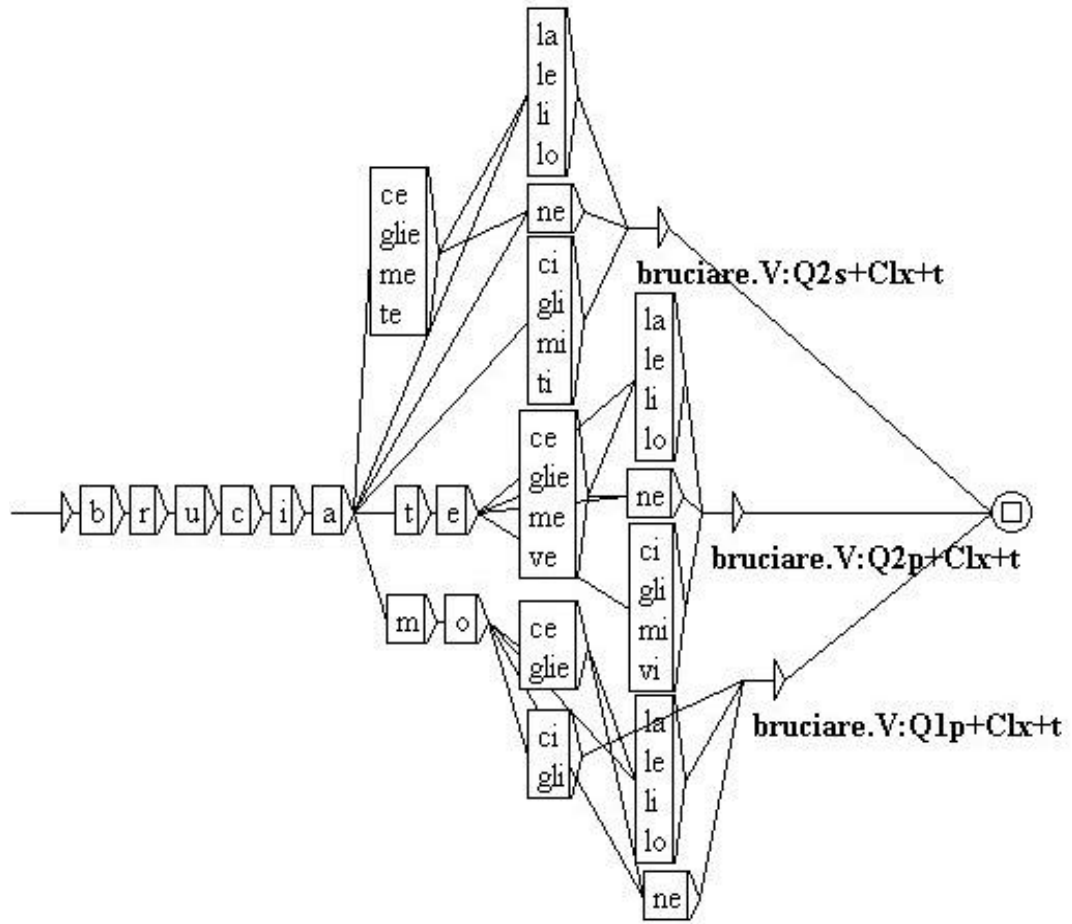


Figure 11: graphe pour l'impératif.

Ce graphe reconnaît et étiquette les agglutinations suivantes (75 formes):

bruciaci, bruciagli, bruciami, bruciati, bruciala, bruciale, bruciali, brucialo, bruciane, bruciacelo, bruciacela, bruciaceli, bruciacele, bruciacene, bruciamelo, bruciamela, bruciameli, bruciamele, bruciamene, bruciatelo, bruciatela, bruciateli, bruciatele, bruciatene,

bruciaglielo, bruciagliela, bruciaglieli, bruciagliele, bruciagliene, bruciatela, bruciatele, bruciateli, bruciatelo, bruciatene, bruciateci, bruciategli, bruciatemi, bruciatevi, bruciatecela, bruciatecele, bruciateceli, bruciatecelo, bruciatecene, bruciategliela, bruciategliele, bruciateglieli, bruciateglielo, bruciategliene, bruciatemela, bruciatemele, bruciatemeli, bruciatemelo, bruciatemene, bruciatevela, bruciatevele, bruciateveli, bruciatevelo, bruciatevene, bruciamolo, bruciamola, bruciamoli, bruciamole, bruciamone, bruciamoci, bruciamogli, bruciamocelo, bruciamocela, bruciamoceli, bruciamocene, bruciamocene, bruciamoglielo, bruciamogliela, bruciamoglieli, bruciamogliele, bruciamogliene

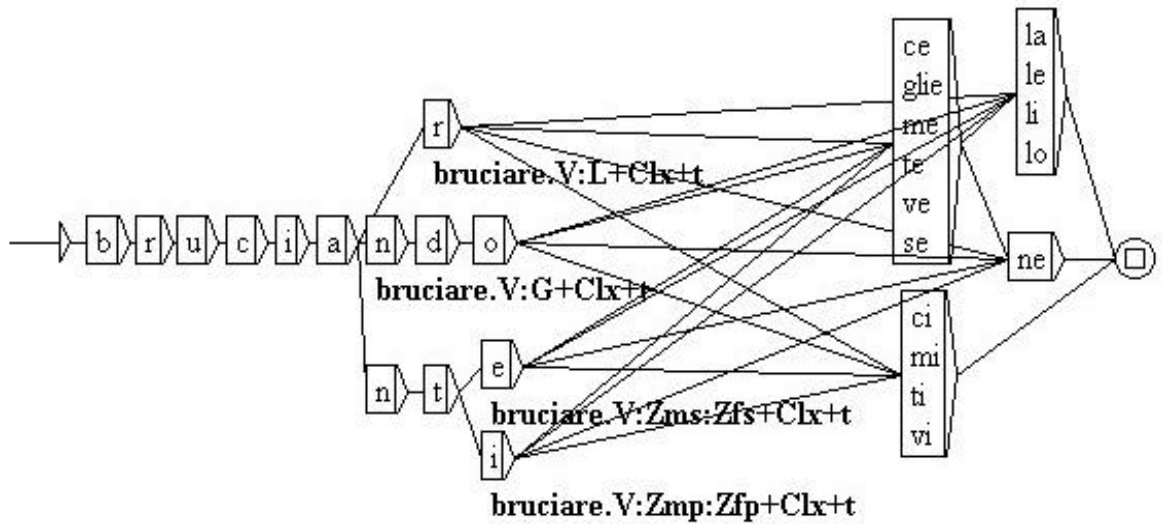


Figure 12: graphe pour l'infinitif, le gérondif et le participe présent.

Ce graphe reconnaît et étiquette les agglutinations suivantes (164 formes):

bruciarlo, bruciarla, bruciarli, bruciarle, bruciarmi, bruciarti, bruciarci, bruciarvi, bruciarli, bruciarne, bruciarmelo, bruciarmela, bruciarmeli, bruciarme, bruciarmene, bruciantelo, bruciantela, brucianteli, bruciantele, bruciantene, bruciarglielo, bruciargliela, bruciarglieli, bruciargliele, bruciargliene, bruciarcelo, bruciarcela, bruciarceli, bruciarcele, bruciarcene, bruciarvene, bruciarvelo, bruciarvela, bruciarveli, bruciarvele, bruciarvene, bruciarselo, bruciarsela, bruciarseli, bruciarsele, bruciarsene, bruciandolo, bruciandola, bruciandoli, bruciandole, bruciandomi, bruciandoti, bruciandoci, bruciandovi, bruciandoli, bruciandone, bruciandomelo, bruciandomela, bruciandomeli, bruciandomele, bruciandomene, bruciandotelo, bruciandotela, bruciandoteli, bruciandotele, bruciandotene, bruciandoglielo, bruciandogliela, bruciandoglieli, bruciandogliele, bruciandogliene, bruciandocelo, bruciandocela, bruciandoceli, bruciandocele, bruciandocene, bruciandoselo, bruciandosela, bruciandoseli, bruciandosele, bruciandosene, bruciandovene, bruciandovelo, bruciandovela, bruciandoveli, bruciandovele, bruciandovene, bruciantelo, bruciantela, brucianteli, bruciantele, bruciantemi, brucianteti, brucianteci, bruciantevi, brucianteli, bruciantene, bruciantemelo, bruciantemela, bruciantemeli, bruciantemele, bruciantemene, bruciantetelo, bruciantetela, brucianteteli, bruciantetele, bruciantetene, brucianteglielo, bruciantegliela, brucianteglieli, bruciantegliele, bruciantegliene, brucianteselo, bruciantesela, brucianteseli, bruciantesele, bruciantesene, bruciantecelo, bruciantecela, brucianteceli, bruciantecele, bruciantecene, bruciantevene, bruciantevelo, bruciantevela, brucianteveli, bruciantevele, bruciantevene, bruciantilo, bruciantila, bruciantili, bruciantile, bruciantimi, bruciantiti, bruciantici, bruciantivi, bruciantili, bruciantine, bruciantimelo, bruciantimela, bruciantimeli, bruciantimele, bruciantimene, bruciantitelo, bruciantitela, bruciantiteli, bruciantitele, bruciantitene, bruciantiglielo, bruciantigliela, bruciantiglieli, bruciantigliele,

bruciantigliene, brucianticelo, brucianticela, brucianticeli, brucianticele, brucianticene, bruciantiselo, bruciantisela, bruciantiseli, bruciantisele, brucianteiene, bruciantivene, bruciantivelo, bruciantivela, bruciantiveli, bruciantivele, bruciativene

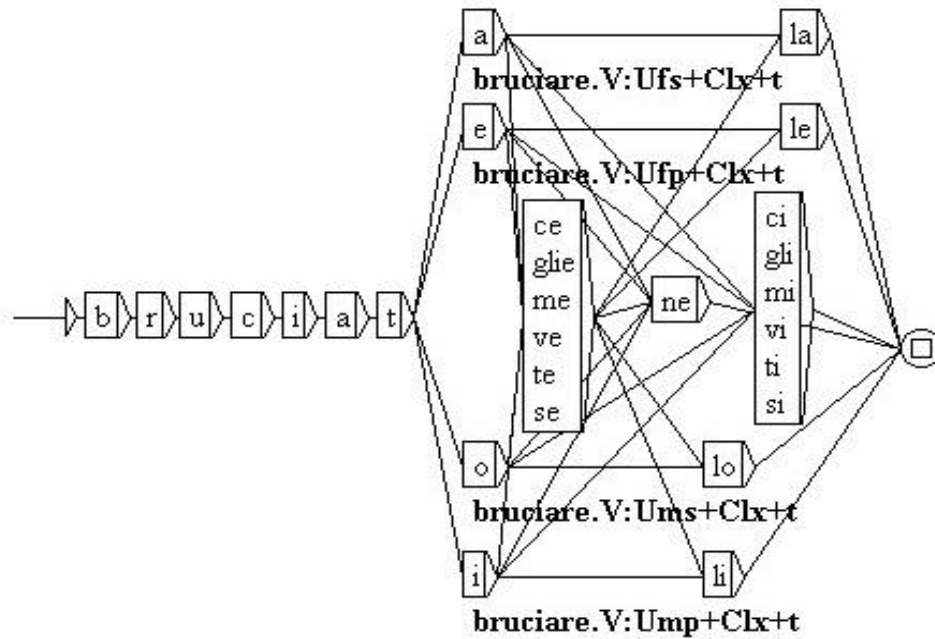


Figura 13: graphe pour le participe passé.

Ce graphe reconnaît et étiquette les agglutinations suivantes (76 formes):

bruciataci, bruciatagli, bruciatami, bruciatavi, bruciatati, bruciatasi, bruciatoci, bruciatogli, bruciatomi, bruciatovi, bruciatoti, bruciatosi, bruciatichi, bruciatigli, bruciatimi, bruciativi, bruciatiti, bruciatisi, bruciateci, bruciategli, bruciatemi, bruciatevi, bruciateti, bruciatesi, bruciatale, bruciatane, bruciatele, bruciatene, bruciatili, bruciatine, bruciatolo, bruciatone, bruciatacela, bruciatagliela, bruciatamela, bruciatavela, bruciatasela, bruciatacene, bruciatagliene, bruciatamene,

*bruciatavene, bruciatatene, bruciatasene, bruciatocelo, bruciatoglielo,
bruciatomelo, bruciatovelo, bruciatoselo, bruciatocene, bruciatogliene,
bruciatomene, bruciatovene, bruciatotene, bruciatosene, bruciatecele,
bruciategliele, bruciatemele, bruciatevele, bruciatesele, bruciatecene,
bruciategliene, bruciatemene, bruciatevene, bruciatetene, bruciatesene,
bruciatriceli, bruciatiglieli, bruciatimeli, bruciativeli, bruciatiseli,
bruciatricene, bruciatigliene, bruciatimene, bruciativene, bruciatitene,
bruciatisene*

3.1.1.4 Quelques considérations sur les verbes pronominaux

Parmi les agglutinations listées nous observons qu'il y en a quelques-unes apparemment non-grammaticales, certes utilisés rarement. Néanmoins, l'analyse syntaxique sur laquelle elles se basent les rend amplement valables, et il suffira de refaire à rebours le parcours de la pronominalisation, en partant d'une agglutination donnée, pour obtenir des phrases plus que grammaticales.

Outre que justifier l'existence et l'usage de formes spécifiques, cette procédure simple et efficace confirme la validité de notre méthode de traitement des agglutinations, non seulement comme approche inférentielle de dérivation, mais aussi comme test possible sur l'usage et la mise à jour du lexique de l'italien, au cas où des doutes à propos de la grammaticalité de quelques mots devraient être vérifiés.

Mais avant d'exposer nos dernières considérations sur ce sujet, nous devons chercher les réponses les plus exhaustives à la question suivante: pourquoi avons-nous besoin de reconnaître les agglutinations de l'italien? La raison, comme on l'a déjà dit, en est dans le fait que les formes agglutinées en italien sont très utilisées et peuvent être appliquées à presque tous les verbes, exception faite pour ceux qui présentent des restrictions sémantiques et syntaxiques évidentes. Les listes d'agglutinations des paragraphes précédents, dans lesquelles nous avons aussi signalé le nombre de formes reconnues par chaque automate, incluent précisément 614⁷¹ séquences. En considérant le fait que nous avons construit et appliqué seulement 12 graphes, nous observons que chaque graphe reconnaît en moyenne 52 formes. L'italien a plus de 12.000 verbes et plus de 30.000 emplois verbaux. Nous pouvons donc supposer que les agglutinations potentielles de la langue italienne sont environ au nombre de 18.420.000, nombre qui confirme leur importance, surtout en termes statistiques.

Toutefois, il y a d'autres raisons pour lesquelles il nous semble nécessaire de reconnaître automatiquement les agglutinations. Dans les dictionnaires papier de

⁷¹ Pour ce calcul, nous n'avons pas compté les homographes, parce que, même si ces mots sont formellement égaux, ils ont néanmoins des significés différents et aussi des dérivations syntaxiques différentes, donc ils sont des formes uniques.

l'italien, beaucoup d'infinitifs pronominaux dont dérivent ces formes ne sont pas lemmatisés, et si ce choix de lemmatisation peut être justifié pour les agglutinations syntaxiques, certes il ne l'est pas pour les agglutinations morphologiques, qui, comme nous avons pu le voir, peuvent, être obtenues seulement en fléchant les formes à l'infinitif. Dans les dictionnaires papier, en général, les verbes pronominaux dérivés morphologiquement de verbes ordinaires sont seulement listés comme usages particuliers dans les gloses de ces derniers; mais il serait plus naturel de les lemmatiser séparément. Cette dernière observation se justifie par le fait qu'à l'intérieur des usages d'un même verbe, on peut trouver de remarquables différences formelles, syntaxiques et sémantiques, comme dans les phrases suivantes, réalisées avec le verbe *accostare* (accoster, approcher):

1. *Max accosta a destra* (Max accoste à droite)
2. *Max accosta la macchina al marciapiede* (Max accoste la voiture au trottoir);
3. *Questo colore si accosta al verde* (Cette couleur va bien avec le vert: intransitif pronominal);
4. *La macchina si accosta al marciapiede* (La voiture s'approche du trottoir: intransitif pronominal)
5. *Max si accosta a Lea* (Max s'approche de Léa: réflexif)

Les usages pronominaux sont signalés de façon plutôt discontinue⁷². Si pour *accostare* nous trouvons en général une description exhaustive dans tous les ouvrages,

⁷² Mais il faut souligner que les classifications syntaxiques sont tout autant hétérogènes. Par exemple, les deux verbes *inginocchiarsi* et *genuflettersi* (s'agenouiller, pour les deux) dans des ouvrages différents sont parfois lemmatisés comme réflexifs, parfois comme intransitifs, et aussi avec des notations innovatrices telles que *verbe réflexif intransitif*.

les notations sont assez différentes pour *calunniare* (calomnier), verbe transitif qui peut avoir deux formes pronominales:

Max e Luca si calunniano (Max et Luc se calomnient: réciproque)

Max si calunnia (Max se calomnie: réflexif)

qui ne sont presque jamais lemmatisées. Un traitement similaire est réservé aux usages transitifs pronominaux, comme par exemple celui du verbe *causarsi* (se causer) comme dans la phrase:

8. *Max si causa un grande dolore* (Max se cause une grande douleur)

qui peut être réécrite comme:

8a. *Max causa un grande dolore a se stesso*
(Max cause une grande douleur à soi même)

Aussi, souvent les emplois verbaux de possession ne sont pas signalés, comme pour *cavarsi* (s'arracher) dans la phrase:

9. *Max si cava un dente* (Max s'arrache une dent)

La typologie des absences est donc plutôt hétérogène mais elle ne peut pas être réduite à un choix spécifique ou déclaré. Aussi, en ce cas, vaut la considération faite

précédemment, i.e. que l'absence et la classification fautive d'un lemme dans un dictionnaire rendent plus difficile sinon improbable son usage et sa compréhension. Si nous considérons de nouveau le nombre indiqué pour les agglutinations potentielles de l'italien, nous pouvons conclure que pour la lexicographie classique le vide lexical qui dérive de ces absences ou classifications fautives est sûrement remarquable.

Heureusement, la linguistique informatique est équipée de moyens *non intelligents* aptes à appliquer récursivement et économiquement, i.e. en les modifiant légèrement, toutes les opérations de lecture automatique que nous avons analysées jusqu'ici et qui nous ont donné des résultats très satisfaisants. Par exemple, en revenant à Intex®, nous observons que l'automate réalisé pour *bruciare*, usage transitif, peut facilement être modifié de façon à être utilisé pour tout verbe transitif qui ait les mêmes propriétés distributionnelles. En outre, si les graphes du verbe *bruciare* peuvent être modifiés et appliqués avec succès pour la reconnaissance des agglutinations d'autres verbes transitifs, et si nous raisonnons en termes elliptiques, il sera possible de construire un seul automate apte à gérer la reconnaissance automatique des agglutinations de tous les verbes transitifs ayant une distribution identique, i.e. d'un point de vue lexico-grammatical il sera possible d'adopter un seul automate pour tous les verbes de la même classe syntaxique.

Par exemple, pour reconnaître les agglutinations à l'infinitif, gérondif et participe présent du verbe *tagliare* (couper) à partir de la phrase:

10. *Max taglia il dolce di Paolo* (Max coupe le gâteau de Paul)

il sera suffisant de substituer dans les nœuds initiaux du graphe de la figure 12, la séquence *taglia-* à la séquence *brucia-*, pour obtenir:

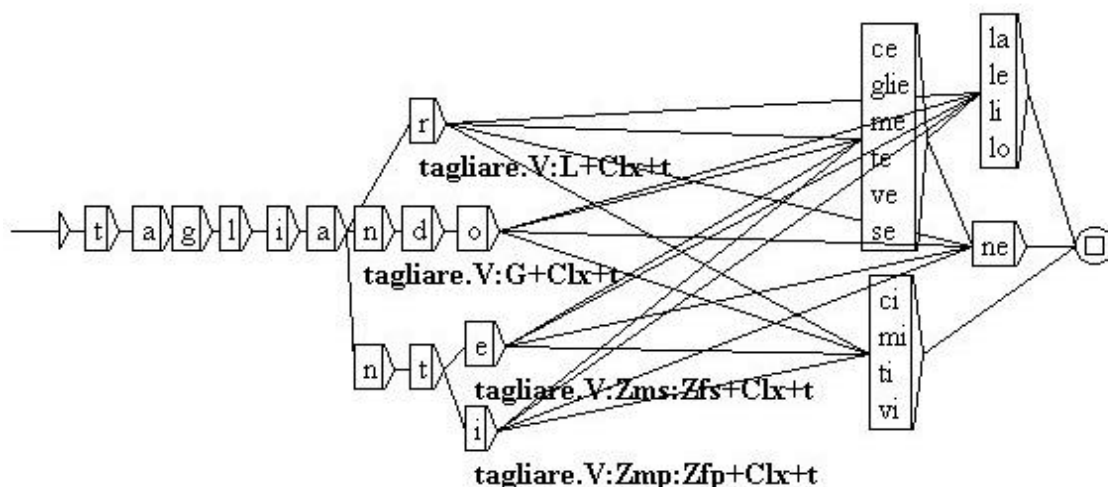


Figure 14: graphe pour l'infinitif, le gérondif et le participe présent du verbe *tagliare* (couper), usage transitif.

Cette opération ne requiert que quelques minutes, et il nous faudrait la même quantité de temps pour adapter à *tagliare* tous les autres graphes réalisés pour *bruciare*. Cela démontre qu'avec Intex® il est possible de créer des automates pour des classes syntaxiques entières, en faisant sorte que chaque classe ait les siens. De cette façon, et en considérant que “(...) la longueur et la complexité des FSA n'ont pas de limites, donc ils peuvent représenter un haut nombre de phénomènes. Les FSA peuvent aussi être en cascade, donc il est possible d'utiliser en même temps un nombre élevé de grammaires...”⁷³, la reconnaissance automatique des formes agglutinées peut aboutir à de hauts niveaux d'exhaustivité et donc représenter une étape fondamentale pour le remplissage de ces vides lexicaux, donc linguistiques aussi, créés par les descriptions fautives ou défectueuses des dictionnaires papier.

⁷³ Silberztein, M. (1996), page 10 (c'est nous qui traduisons).

3.2. LA GUERRE ENTRE LES LEXIQUES

Après avoir étudié les propriétés de lemmes spécifiques, il peut arriver que l'on choisisse de leur donner dans les dictionnaires électroniques des classifications grammaticales différentes de celles fournies par les dictionnaires papier, en générant ce que l'on peut appeler *la guerre entre les lexiques*. Il peut se vérifier que, dans les deux dictionnaires, une même unité lexicale se présente avec des descriptions grammaticales assez hétérogènes, parfois inconciliables même si apparemment légitimées par des analyses soignées et aussi par la littérature.

Nous avons déjà dit que pour classer correctement les unités lexicales d'une langue il est nécessaire d'en étudier taxinomiquement tous les usages possibles, dans la langue parlée ou dans la langue écrite, et aussi baser leur observation sur un modèle d'analyse fort et consolidé, comme par exemple celui fourni par le transformationalisme ou le lexique-grammaire. Notre choix d'adopter ce dernier, provient du fait que nous réputons le lexique-grammaire la méthode la plus adéquate pour résoudre les problèmes qui concernent les ambiguïtés et la classification des unités lexicales. C'est justement grâce au lexique-grammaire que nous avons repéré quelques-unes des particularités de l'italien, dont nous n'avions pas connaissance auparavant et que nous exposerons dans les pages suivantes. Comme toutes les guerres celle-ci aussi se terminera sans vainqueurs et vaincus, mais elle aidera probablement à mettre en évidence des éléments morpho-syntaxiques de l'italien qui sont souvent négligés ou considérés comme acquis.

3.2.1 Pronoms ou numéraux?

Par définition, un pronom est un élément utilisé ou utilisable en lieu d'un substantif, dont il reprend les mêmes valeurs grammaticales, de genre et souvent de nombre. La fonction principale du pronom est donc celle de se substituer à un nom, de renvoyer à celui-ci, donc d'être avec celui-ci dans un rapport de phoracité, dans la même phrase ou à son extérieur, comme dans l'exemple qui suit:

1. *Arrivarono le **macchine**. Max mise in moto la **sua**, poi la **tua** ed infine la **mia***

(Les voitures arrivèrent. Max mit en marche la sienne, puis la tienne et enfin la mienne)

Dans la phrase (1), les mots en gras peuvent être séparés en deux groupes, le premier composé par **macchine**, le deuxième par **sua**, **tua**, **mia**, qui avec **macchine** sont en rapport d'anaphore. Les mots du deuxième groupe sont donc tous des pronoms au singulier utilisés en référence avec un substantif pluriel ou bien avec les unités simples auxquelles le substantif **macchine** fait référence. Ce processus est aussi plus évident dans le discours suivant, où le référent anaphorique est expressément cité:

2. *Arrivarono le **macchine**. Max mise in moto la **sua macchina**, poi la **tua macchina** ed infine la **mia macchina***

(Les voitures arrivèrent. Max mit en marche sa voiture, puis ta voiture et enfin ma voiture)

D'un point de vu formel et syntaxique, nous observons que l'itération du substantif **macchina** en (2) transforme en adjectifs (i.e. dans le sens de la grammaire italienne classique) les termes qui en (1) étaient au contraire des pronoms possessifs.

Considérons par contre ces deux phrases:

3. *Le **atlete** cominciarono ad arrivare al traguardo. Ida fu la **terza***

(Les athlètes commencèrent à parvenir à la ligne d'arrivée. Ida fut la troisième)

4. *I **gelati** erano sul bancone. Max mangiò il **secondo***

(Les glaces étaient sur le comptoir. Max mangea la deuxième)

ou:

- 3a. *Ida fu la **terza** ad arrivare al traguardo*
(Ida fut la troisième à parvenir à la ligne d'arrivée)

- 4a. *Max mangiò il **secondo** dei gelati che erano sul bancone*
(Max mangea la deuxième des glaces qui étaient sur le comptoir)

Nous observons qu'en (3) et en (3a) **terza**, même s'il s'agit d'un adjectif numéral ordinal, a avec **atlete** un rapport phorique typiquement pronominal, et on peut dire la même chose pour **secondo** e **gelati** des phrases (4) et (4a). Seulement dans (3a) le terme de phoricité est sous-entendu, et le contexte porte indirectement à une anaphore avec **atlete**, i.e. à une interprétation pronominale.

Nous nous trouvons donc devant l'hypothèse que les adjectifs numéraux ordinaux aient aussi une valeur pronominale. Outre que par la phoricité des éléments linguistiques, cette hypothèse peut être vérifiée grâce à d'autres analyses, comme par exemple l'analyse harrisienne de l'équivalence distributionnelle entre des couples de phrase tel que:

5. *Fra le macchine disponibili, Max ha preso la **mia***
(Parmi les voitures disponibles, Max a pris la mienne)

6. *Fra le macchine disponibili, Max ha preso la **seconda***
(Parmi les voitures disponibles, Max a pris la deuxième)

Nous observons que **seconda** en (6) est en rapport d'équivalence distributionnelle avec le pronom **mia** de (5). Cette équivalence donne à **seconda** les caractéristiques de pronom, en confirmant donc l'hypothèse que les numéraux pronominaux sont des pronoms.

Le même type d'analyse est applicable pour les adjectifs numéraux cardinaux, comme le démontrent les phrases suivantes:

7. *Arrivarono le **macchine**. Max prese le **due** che gli erano state promesse*
(Les voitures arrivèrent. Max prit les deux qu'on lui avait promis)

8. *I **panini** erano pronti. Max mangiò **uno** dei più morbidi*
(Les sandwiches étaient prêts. Max prit un des plus moelleux)

En ce cas aussi, en ajoutant à droite des deux pronoms les référents phoriques, nous aurons des adjectifs:

- 7a. *Arrivarono le **macchine**. Max prese le **due macchine** che gli erano state promesse*
(Les voitures arrivèrent. Max prit les deux voitures qu'on lui avait promis)

- 8a. *I **panini** erano pronti. Max mangiò **un panino** dei più morbidi*
(Les sandwiches étaient prêts. Max prit un sandwich des plus moelleux)

Sur ces bases, dans le DELAS et donc aussi dans le DELAF, les adjectifs ordinaux et cardinaux ont été classifiés par ailleurs comme pronoms. Même si l'analyse précédente ne laisse que peu de doutes, cette classification n'est repérable dans aucun dictionnaire papier ou aucune grammaire actuellement dans le commerce et que nous avons consulté, pour l'italien et pour d'autres langues, dans lesquels ces éléments sont étiquetés seulement comme substantifs et adjectifs. Seule exception est l'*Oxford Advanced Learner's Dictionary*, édité par A. S. Hornby, dans lequel tous les numéraux et les ordinaux sont étiquetés aussi comme pronoms et non comme substantifs. Nous soulignons que d'autres unités lexicales semblent pouvoir figurer dans des contextes tout à fait similaires à ceux que nous avons vus pour les numéraux et les

ordinaux. La situation est donc plutôt complexe et elle mériterait des approfondissements ultérieurs, comme le démontre la phrase suivante:

9. *Ida fu (l'ultima + la penultima + la terzultima + l'ennesima + la sola + l'unica) ad arrivare al traguardo*
(Ida fut (la dernière + l'avant-dernière + la troisième avant la dernière + l'énème + la seule + l'unique) à parvenir à la ligne d'arrivée)

3.2.2 Countables et Uncountables

Pour quelques classes de substantifs, lors de la création du DELAS et du DELAF la distinction entre emplois nombrables et non nombrables a été appliquée. C'est une distinction typique de la lexicographie anglo-saxonne, qui utilise les notations de *countable* et *uncountable*. Des ces notations, par contre, il n'y a aucune trace dans les dictionnaires italiens papier qui sont dans le commerce. Comme nous le verrons, la différenciation entre *countable* et *uncountable* est logique et linguistique aussi; elle est très importante en termes morpho-grammaticaux, sémantiques est syntaxiques, et elle est aussi très utiles pour que des FSA reconnaissent de particulières séquences de texte de façon non ambiguë.

Les classes de substantifs pouvant être catalogués soit comme nombrables soit comme non nombrables sont les suivants:

- a) la classe des noms de certaines matières, comme par exemple *legno* (bois), *plastica* (plastique) ou *vetro* (verre). L'usage au singulier de ces mots est presque une norme dans des phrases comme:

1. *Questa casa è in legno*
(Cette maison est en bois)
2. *La bottiglia è ricoperta di plastica*
(La bouteille est recouverte de plastique)
3. *Il parabrezza è fatto di vetro*
(Le pare-brise est fait en verre)

Si dans ces phrases nous essayons de substituer les formes au pluriel, nous obtenons:

1a. **?Questa casa è in legni*

(Cette maison est en bois)

2a. *La bottiglia è ricoperta di plastiche*

(La bouteille est recouverte de plastiques)

3a. **?Il parabrezza è fatto di vetri*

(Le pare-brise est fait en verres)

i.e. deux phrases grammaticalement inacceptables et difficilement compréhensibles (1a et 3a), et une phrase avec un signifié différent de celui originel (2a). Néanmoins, les phrases:

4. *Il portiere è stato superato dalla palla ma un legno della sua porta lo ha salvato*

(Le gardien de but a été dépassé par le ballon mais un des poteaux du but l'a sauvé)

c'est-à-dire:

5. *Il portiere è stato superato dalla palla ma i legni della sua porta lo hanno salvato*

(Le gardien de but a été dépassé par le ballon mais les poteaux du but l'ont sauvé)

sont au contraire des phrases grammaticales dans lesquelles, par rapport à (1a), on donne à *legno* non pas le signifié de part solide du fût des rameaux et des racines des arbres et des arbustes, mais le signifié idiomatique d'objet réalisé avec cette partie solide et utilisé pour construire les buts du football⁷⁴. Cela nous porte à reconnaître l'existence d'un substantif masculin *legno* qui est seulement au singulier et non nombrable, et un substantif masculin *legno* avec une forme au pluriel *legni*, donc nombrable;

b) la classe des noms des éléments chimiques comme *argento* (argent), *ferro* (fer) et *oro* (or), pour les mêmes raisons que *legno*. Il semblerait exister une différence remarquable entre l'élément chimique, i.e. la matière de *una sbarra d'oro* (une barre en or) ou *una piastra d'oro* (une plaque en or) et les usages métonymiques en relation avec la matière même, comme ceux existant dans *i ferri del mestiere* (les outils) ou *gli ori di Taranto* (les ors de Tarente). Pour confirmation de cela, nous soulignons que dans la table de Mendeleïev, les éléments chimiques sont classés et se différencient sur la base de caractéristiques et de propriétés spécifiques, en mettant en évidence le fait que chacun d'entre eux est unique et simple, i.e. non nombrable: il existerait donc seulement le *niobio* (niobium) et non les *niobii* (les niobiums). Cela semble vraisemblable aussi pour d'autres éléments tels que les isotopes, parce que nous savons qu'un seul facteur distinctif est suffisant, en relation avec le poids atomique ou d'autres caractéristiques, pour différencier par exemple le *deuterio* (deutérium) de l'*idrogeno* (hydrogène). Une même conclusion peut en être tirée pour les substances chimiques, les protéines et en général pour tout ce qui peut être transcrit à l'aide d'une formule. A ce propos, nous soulignons que le rapport qui existe entre les formules et les entités qu'elles dénotent est univoque, i.e. une formule peut indiquer et dénoter une seule substance. Changer un seul caractère dans la transcription d'une formule quelconque veut dire indiquer une substance nouvelle, différente de celle indiquée par la formule originelle;

⁷⁴ En outre, nous savons aussi que souvent les poteaux de ces buts ne sont même pas réalisés en bois, mais en d'autres matériaux.

c) les noms de gibier et de boucherie. Dans des phrases telles que:

6. *Oggi ho mangiato del vitello*
(Aujourd'hui j'ai mangé du veau)

et

7. *Oggi ho visto un vitello*
(Aujourd'hui j'ai vu un veau)

les signifiants du mot *vitello* indiquent des entités et des objets différents, puisque le premier est un usage elliptique pour *carne di vitello* (chair de veau), tandis que le deuxième est utilisé en référence à un quadrupède ruminant. D'ailleurs, les signifié de phrases comme:

8. *Oggi ho mangiato dei vitelli*
(Aujourd'hui j'ai mangé des veaux)

dans lesquelles *vitelli* ne peut pas être interprété comme *carne di vitello*, semble confirmer l'existence d'un lemme *vitello* utilisable seulement au singulier et d'un autre lemme *vitello* avec une forme plurielle.

Comme nous l'avons déjà dit, les notations *numerabile* et *non numerabile* ne sont pas utilisées dans les dictionnaires papier de l'italien, tandis que pour les substantifs

comme ceux que nous venons d'analyser, dans le DELAS et le DELAF nous avons prévu des étiquettes aptes à différencier les emplois seulement au singulier de ceux au singulier et au pluriel. En reprenant une classification du lexique-grammaire, dans les dictionnaires électroniques les substantifs *uncontables* ont été définis comme *nomi massa* (noms de masse). L'application de ce type de descriptions rend les listes de lemmes électroniques plus détaillées et en outre, pendant la construction des FSA d'Intex®, elle permet d'utiliser des instructions morphologiques très précises qui par conséquent apportent des résultats non ambigus et hautement crédibles. En fait, puisque, dans les phrases simples et les discours, les noms de masse peuvent figurer seulement au singulier, ils imposent des restrictions de nombre à tous les mots avec lesquels ils entrent en rapport de co-référence. Grâce à cette observation et à la différenciation entre étiquettes nombrables et non nombrables, pendant l'analyse textuelle avec Intex® il sera donc possible de reconnaître automatiquement les différents usages lexicaux de noms spécifiques, et aussi les différents signifiés de phrases telles que:

Il cristallo di questo bicchiere è bellissimo

(Le cristal de ce verre est très beau)

I cristalli di Max sono bellissimi

(Les glaces de Max sont très belles)

3.2.3 Le lexique en fonction du lexique

L'étiquette *in funzione di* (en fonction de) est souvent utilisée dans les dictionnaires papier pour ces lemmes qui, appartenant à une catégorie grammaticale donnée, peuvent être utilisés par substitution à d'autres lemmes, de catégories différentes.

Par exemple, dans:

1. *Max camminava lento verso casa*
(Max marchait lentement vers la maison)

nous observons que la forme *lento* (lent) est utilisé avec une fonction adverbiale, puisque:

- d'un point de vue distributionnel, peut se substituer à l'adverbe *lentamente* (lentement), dont elle reprend le signifié et avec lequel celui-ci est en rapport de dérivation morphologique, comme le démontre la phrase:

- 1a. *Max camminava lentamente verso casa*
(Max marchait lentement vers la maison)

- elle répond à une question commençant par *come* (comment), qui peut être utilisée comme test formel pour la définition des adverbes de "manière"⁷⁵:

⁷⁵Voir Gross, M., 1991.

1b. *Come camminava Max verso casa? - Lento*

(Comment marchait Max vers la maison? - Lentement)

- il représente un complément de phrase⁷⁶, i.e. il n'est pas un complément essentiel du verbe *camminare*, parce qu'il est possible de l'effacer sans modifier la grammaticalité et le sens de la phrase même:

1c. *Max camminava verso casa*

(Max marchait vers la maison)

Néanmoins, nous ajoutons que, même si la catégorie grammaticale de l'adverbe ne prévoit pas de formes fléchies, la phrase suivante montre comment les adjectifs utilisés en fonction adverbiale gardent leurs caractéristiques morphologiques flexionnelles:

1. *Ida camminava lenta verso la casa*

(Ida marchait lentement vers la maison)

tandis que nous aurons:

2. **Ida camminava lento verso la casa*

⁷⁶Voir Annibale Elia, Martinelli, M., D'Agostino, E., 1981.

Dans les dictionnaires papier de l'italien, cette fonction possible comme adverbe n'est pas signalée systématiquement, comme il se vérifie par exemple avec *feroce* (féroce), qui dans la phrase:

3. *Il cane abbaiò feroce al ladro*
(Le chien aboya féroce après le voleur)

a la même fonction adverbiale que *lento* en (3). La même chose peut être dite au sujet de *veloce*, qui est utilisé adverbialement dans des phrases telles que:

4. *L'attaccante corre veloce verso l'area avversaria*
(L'avant court rapidement vers la surface adverse)

En italien, les verbes aussi peuvent être utilisés dans une autre fonction, plus précisément celle de substantif, comme il est possible de le vérifier dans:

5. *Il tuo (parlare + urlare + dormire + chiacchierare) snerva Max*
(Le fait que tu (parles + hurles + dors + bavardes) énerve Max)

ou aussi en:

6. *Luca si accorge (del parlare + dell'urlare + del dormire + del chiacchierare) di Max*
(Luc s'aperçoit du fait que Max (parle + hurle + dort + bavarde))

Il est néanmoins important de souligner que ces emplois présentent des différences remarquables par rapport à celles que nous avons vues pour les adverbiaux adjectivaux, parce que les verbes qui figurent en fonction nominale:

- d'un point de vue distributionnel, ne peuvent pas toujours être remplacés par des éléments qui ont le même sens, une catégorie grammaticale différente et sont en rapport de dérivation morphologique avec ceux-ci. Pour les phrases (6) et (7), la substitution produit des résultats grammaticaux seulement dans deux cas, c'est-à-dire:

6a. *Il tuo (urlo + chiacchierio) snerva Max*
(Ton (hurlement + babillage) énerve Max)

7a. *Luca si accorge (dell'urlo + del chiacchierio) di Max*
(Luc s'aperçoit (du (hurlement + babillage) de Max)

- sont des compléments essentiels du verbe qui les sélectionne, et il n'est pas possible de les effacer sans modifier le signifié, la syntaxe et la grammaticalité des phrases dans lesquelles ils figurent:

6b. **Il tuo snerva Max*
(*Ton énerve Max)

7b. **Luca si accorge del di Max*
(*Luc s'aperçoit du de Max)

- d'un point de vue morphologique, ils ne prennent pas les caractéristiques du substantif, parce qu'ils n'ont aucune flexion au pluriel, ni aucun emploi au pluriel:

6c. **?I tuoi (**parlare** + **urlare** + **dormire** + **chiacchierare**) snervano Max*

7c. *I tuoi (***parlari** + ***urlari** + ***dormiri** + ***chiacchierari**) snervano Max*

- d'un point de vue transformationnel, leurs occurrences en fonction nominale sont des réductions de phrases complétives sujet et objet, comme il est possible de le vérifier dans les phrases suivantes:

6. *Il tuo (**parlare** + **urlare** + **dormire** + **chiacchierare**) snerva Max*

[T Compl Sogg] =:

6d. *Il fatto che tu (**parli** + **urli** + **dorma** + **chiacchieri**) snerva Max*

(Le fait que tu (parles + hurles + dors + bavardes) énerve Max)

7. *Luca si accorge (**del parlare** + **dell'urlare** + **del dormire** + **del chiacchierare**) di Max*

[T Compl Ogg] =:

7e. *Luca si accorge del fatto che Max (**parla** + **urla** + **dorme** + **chiacchiera**)*

(Luc s'aperçoit du fait que Max (parle + hurle + dort + bavarde))

et dépendent de la syntaxe de l'élément prédicatif qui les sélectionne, i.e. pour les exemples précédents de celle de *snervare* et d'*accorgersi*. En fait, en substituant les éléments prédicatifs de (6d) et (7d) avec deux verbes qui ne sélectionnent pas de complétives sujet et/ou objet, par exemple *giudicare* (juger) et *partire* (partir), nous obtenons des phrases non grammaticales:

6f. **Il fatto che tu (parli + urli + dorma + chiacchieri) giudica Max*
(*Le fait que tu (parles + hurles + dors + bavardes) juge Max)

7f. **Luca parte il fatto che Max (parli + urli + dorma + chiacchieri)*
(*Luc part le fait que Max (parle + hurle + dort + bavarde))

Dans le DELAS et le DELAF, l'usage nominal possible des verbes n'a pas été signalé, parce que ce phénomène, qui comme nous l'avons vu, dépend des transformations syntaxiques, ne peut trouver place dans un dictionnaire électronique morpho-grammatical. Néanmoins, pendant la phase de parsing, il est possible de créer des FSA avec Intex® qui reconnaissent quand un verbe figure en fonction de complément nominal dans une phrase à complétive sujet et/ou objet

Par contre, pour revenir au sujet des dictionnaires papier de l'italien, nous observons que l'étiquette *in funzione di* est utilisée sporadiquement, même si presque tous les verbes italiens peuvent être utilisés comme noms, comme le démontre la phrase suivante:

7. *Uccidere non può mai essere soggetto di un verbo*
(Tuer ne peut jamais être sujet d'un verbe)

Dans les dictionnaires papier, cette notation est appliquée tant à des verbes non-fréquemment utilisés, comme par exemple *tremolare* (trembler), qu'à des verbes de haute fréquence, i.e. *peggiore* (aggraver) ou *piangere* (pleurer), et en général elle prend une valeur ni précise ni systématique. Il n'est donc possible d'affirmer que l'application de l'étiquette *in funzione di* constitue une méthode descriptive rigoureuse. Il est aussi à souligner qu'elle est adoptée pour indiquer des phénomènes assez différents entre eux. En fait, pour les adjectifs utilisés en fonction d'adverbe, la transcatégorisation fonctionnelle est un fait concret, tandis que l'utilisation de verbes en fonction de substantifs est possible seulement à partir d'un élément prédicatif qui permet la réduction d'une complétive. Si dans le premier cas, l'adjectif est en relation stricte avec l'adverbe auquel il se substitue, dans le deuxième cas nous nous trouvons vis-à-vis d'une paraphrase syntaxique, dans laquelle il s'agit du syntagme nominal (et il fonctionne comme substantif) et non du verbe.

POUR NE PAS CONCLURE

Nous sommes donc arrivés à la fin de notre discours sur la langue italienne et sur son patrimoine lexical, et il est donc nécessaire de tirer des conclusions à partir de l'ensemble de ce que nous avons dit. D'emblée, il semble que notre approche impose un nouveau point de vue sur la représentation du lexique italien. Même si notre but initial était d'illustrer la genèse et les fonctions d'un dictionnaire électronique, et aussi les différences avec un dictionnaire papier, dans les pages précédentes nous aussi avons touché à des questions non directement pertinentes par rapport à la lexicographie et à la linguistique informatique, mais qui en tout cas se développent à partir de leur relation. Nous avons donc essayé de répondre à des questions différentes et importantes; néanmoins, parvenus à la ligne d'arrivée, nous nous trouvons vis-à-vis d'autres questions, elles aussi d'une importance cruciale. Par exemple, en considérant une période historique, quel type de rapport existe entre les dictionnaires d'une langue, qui sont construits sur un corpus, donc sur la *parole* saussurienne, et son lexique potentiel, i.e. ce système non parfait mais perfectible, qui inclut un nombre de mots beaucoup plus nombreux que celui qui peut être extrait par un corpus et représenté dans un dictionnaire papier? Et puis, quelle justification théorique nous pouvons donner à propos du choix de lemmatiser dans les dictionnaires électroniques tous les termes doués de dicibilité et sémantiquement prégnants?

D'un point de vue strictement sémiologique, nous pouvons peut-être répondre aux questions précédentes en analysant la production et la compréhension des énoncés tels que:

*Ho visitato la mostra documentaria sull'**eremitismo** nel Salernitano*

(J'ai visité l'exposition sur l'**érémisme** dans la région de Salerne)

et:

Ho seguito le fasi del ***costruttaggio** del nuovo stadio
(J'ai suivi les phases du ***constructage** du nouveau stade)

Dans ces deux phrases, les termes en gras ont été tous les deux formés en respectant les règles morphologiques de l'italien (et du français aussi), et néanmoins nous sommes plus enclins à accepter l'usage d'**eremitismo** et à refuser celui de **costruttaggio**, et nous faisons la même chose pour les deux traductions françaises. Bref, nous acceptons le signifié d'**eremitismo** en relation avec **eremita** (ermite) et refusons celui de **costruttaggio** même s'il est en relation avec **costruire** (construire). Nous lemmatiserons seulement **eremitismo** dans notre dictionnaire électronique, mais affirmer que la norme linguistique empêche l'usage de **costruttaggio** ne sera pas suffisant pour justifier notre choix.

Pinker⁷⁷ cherche à donner une réponse à toutes ces questions, et il soutient que les règles de formation des mots, et aussi celle de production des sons, en général sont génétiquement mémorisées quelque part dans notre cerveau et ne dépendent pas des éléments simples, c'est-à-dire des morphèmes, des phonèmes, des racines et de désinences que nous mettons ensemble en parlant et en écrivant. Il ajoute aussi qu'un facteur efficace de discrimination, à un niveau soit pratique soit théorique, pourrait être aussi ce qui suit: à ce qu'il semble, le lexique est un système complexe – dans le sens des sciences exactes – dont les éléments peuvent avoir les uns avec les autres des relations de type différent, allant de l'interdépendance à l'opposition et à l'inclusion. En outre, le lexique même semble être fondé sur un principe d'économie qui tend à ne pas légitimer la naissance et l'usage d'éléments qui, pendant une période historique donnée, pourraient couvrir des espaces sémiologiques déjà occupés. Ce principe d'économie, qui ne semble pas être *a priori* dans le lexique, formerait la partie centrale du concept de dicibilité, qui permet à un locuteur de langue maternelle de considérer comme prononçables et utilisables, ou de rejeter des mots rencontrés pour la première fois. Pour un locuteur italien, **costruttaggio** ne serait donc pas doué de dicibilité, et son usage serait bloqué par la forme pré-existante **costruzione** (construction), qui est le résultat de

⁷⁷ Voir Pinker, S., ouvrage cité.

procédures philologiques précises et qui a la même racine *costru-* que *costruttaggio*. A son tour, *eremitismo* serait dicible parce qu'il irait se positionner dans un compartiment du lexique sémiologiquement vide, si nous considérons l'absence d'un substantif ayant le même signifié et la même racine. Le principe d'économie justifierait aussi le fait que deux mots différents, dérivés d'une même racine, comme *costrutto* (sens) et *costruzione* (construction), ne peuvent pas être complètement synonymes, ou bien ils peuvent l'être en termes contextuels ou d'occurrence, mais non en termes absolus.

Donc, il est peut-être licite d'affirmer que les dictionnaires papier n'interviennent pas dans la direction du patrimoine lexical potentiel, celui de la *langue* ou de la *compétence*, et qu'une représentation plus complète du lexique se fait en anticipant ce qui est « dicible », au moins pour ce qui concerne les formes de type dérivationnel. En outre, si les grammaires génératives sont des grammaires de la *compétence*, parce qu'elles utilisent le locuteur comme source de données en le privilégiant par rapport aux corpus, donc il est légitime d'utiliser cette même méthode pour étudier à fond la composante lexicale de la langue. En fait, si dans l'enquête linguistique une épreuve cruciale consiste à soumettre aux locuteurs de langue maternelle des énoncés qui dans la réalité existent très rarement, le test sur les mots simples jamais entendus pourra avoir aussi une grande valeur. La société contemporaine, qui, depuis plusieurs décennies, est sur le chemin qui porte à la globalisation, et plus récemment au multimédia, assiste à une prolifération textuelle inconcevable pour les époques précédentes. Les résultats en sont que chaque jour, dans des lectures ou des discours, nous rencontrons des mots nouveaux non lemmatisés dans les dictionnaires et qui néanmoins nous communiquent des choses assez précises. Ils appartiennent donc à ce secteur de la compétence à laquelle les auteurs cherchent à parvenir continuellement. Dans de telles circonstances, la méthode du corpus, telle qu'elle a été appliquée jusque là surtout par la lexicographie classique, n'obtient que des résultats forcément partiels.

En conclusion, nous sentons l'exigence de représenter un lexique qui ait été étudié de façon expérimentale et introspective, une étude que nous chercherons à compléter à l'avenir. Une telle représentation permettrait de donner une solution aux larges trames descriptives des dictionnaires papier, que nous pouvons désormais définir comme un petit miroir avec lequel on voudrait décrire la matière lexicale d'une langue.

Une communauté culturellement évoluée a besoin de normes conventionnelles, certes non-imposées mais qui puissent développer un débat toujours ouvert, dans lequel on veut éviter la position de quelques anglophones, synthétisée par l'expression *I say it so it's English*, en ne prenant pas en considération les inventions spectaculaires et hors du temps, très fréquentes chez nous, et les néologismes non nécessaires, commerciaux ou pseudo-techniques, comme ceux déjà mentionnés *puliastina* et *inputazione*. En ce sens, les normes conventionnelles seraient un moyen pour le lexique de rendre plus riche le patrimoine lexical et culturel, là où, au contraire, d'autres normes, avec des projets descriptifs du même type, comme l'étude du corpus, ont produit des résultats incomplets.

BIBLIOGRAPHIE

AA. VV.

1984: “Grammaire et lexique”, *Revue québécoise de linguistique*, vol. 13. n. 2, Université du Québec à Montréal, Montréal.

1988: *La nuova enciclopedia garzanti delle scienze*, Garzanti, Milano.

1989: *Annales des télécommunications*, 44, n. 1-2, CNET, Issy-les-Moulineaux/Lannion.

1994: *Dizionario della lingua italiana*, Garzanti, Milano.

Baldini, M.

1995: *Storia della comunicazione*, Newton Compton Editori, Roma.

Barcellona, N., Marini, A., Monti, P., Vercesi, M.

1988: *5000 termini dell'informatica*, Gruppo Editoriale Jackson, Milano.

Battaglia, S., Pernicone, V.

1951: *La grammatica italiana*, Loescher, Torino.

Beccaria, G. L., diretto da

1994: *Dizionario di linguistica*, Einaudi, Torino.

Benveniste, E.

1966: *Problèmes de linguistique générale*, Gallimard, Paris (trad. it. *Problemi di linguistica generale*, Il Saggiatore, Milano, 1971).

1969-70: *Le vocabulaire des institutions indo-européennes*, Éditions de Minuit, Paris (trad. it. *Il vocabolario delle istituzioni indoeuropee*, Einaudi, Torino, 1976).

Bloomfield, L.

1933: *Language*, Holt, New York.

Boons, J.-P., Guillet, A., Leclère Ch.

1976: *La structure des phrases simples en français. Constructions intransitives*, Droz, Genève.

Canepari, L.

1979: *Introduzione alla fonetica*, Einaudi, Torino.

Chomsky, N. A.

1965: *Aspects of the Theory of Syntax*, MIT Press, Cambridge, Mass (trad. fr. *Aspects de la théorie syntaxique*, Le Seuil, Paris, 1971).

Chomsky, N. A., Halle, M.

1969: *The Sound Pattern of English*, Harper and Row, New York (trad. fr. *Principes de phonologie générative*, Le Seuil, Paris, 1973).

Courtois, B.

1990: *Dictionnaire électronique du LADL pour les mots simples du français DELAS V06/2*, LADL-CNRS, Université Paris 7, Paris.

Courtois, B., Silberztein, M.

(éds.) 1990: "Dictionnaires électroniques du français", *Langages*, n. 87, Larousse, Paris.

Danlos L.

1985: *Génération automatique de textes en langues naturelles*, Masson, Paris.

(éd.) 1988: "Les expressions figées", *Langages*, n. 90, Larousse, Paris.

Dardano, M.

1978: *La formazione delle parole nell'italiano d'oggi*, Bulzoni, Roma.

Dardano, M., Trifone, P.

1985: *La lingua italiana*, Zanichelli, Bologna.

de' Rossi, B.

1612: *Vocabolario degli Accademici della Crusca*, Giovanni Alberti, Venezia (riedito in *Vocabolario degli Accademici della Crusca*, Le Lettere, Firenze, 1987, presentazione di Giovanni Nencioni).

De Bueriis G.

1992: "Il verbo supporto *avere* con i nomi di malattia" in D'Agostino, E. (a cura di), *Studi di Lessico-Grammatica delle lingue europee*, Loffredo, Napoli.

1995: "Una nota sulla classe dei nomi di grandezze misurabili" in D'Agostino, E. (a cura di), *Tra sintassi e semantica - Descrizioni e metodi di elaborazione automatica della lingua d'uso*, ESI, Napoli.

2003: *Le parole come ordine del mondo*, Editoriale Scientifica, Napoli.

De Bueriis G. et Alii

1992: "La relazione tra morfologia e sintassi: analisi in termini di equivalenze parafrastiche", in D'Agostino, E. (a cura di), *Studi di Lessico-Grammatica delle lingue europee*, Loffredo, Napoli.

1995: "Usi della parola *classe*: operatore elementare e operatore su discorso", in D'Agostino, E. (a cura di), *Tra sintassi e semantica - Descrizioni e metodi di elaborazione automatica della lingua d'uso*, ESI, Napoli.

De Mauro, T.

1980: *Guida all'uso delle parole*, Editori Riuniti, Roma.

Devoto, G., Oli G. C.

1978: *Vocabolario illustrato della lingua italiana*, Le Monnier, Firenze.

1995: *Il dizionario della lingua italiana*, Le Monnier, Firenze.

Dogliotti, M., Rosiello, L., (a cura di)

1995: *Lo Zingarelli 1995*, Zanichelli, Bologna.

1997: *Lo Zingarelli 1997*, Zanichelli, Bologna.

2000: *Lo Zingarelli 2001*, Zanichelli, Bologna.

2001: *Lo Zingarelli 2002*, Zanichelli, Bologna.

2003: *Lo Zingarelli 2004*, Zanichelli, Bologna.

Dubois, J. et Alii

1973: *Dictionnaire de linguistique*, Larousse, Paris (trad. it. *Dizionario di linguistica*, Zanichelli, Bologna, 1979).

Ducrot, O., Todorov, T.

1972: *Dictionnaire encyclopédique des sciences du langage*, Le Seuil, Paris.

Eco, U.

1975: *Trattato di semiotica generale*, Bompiani, Milano.

1984: *Semiotica e filosofia del linguaggio*, Einaudi, Torino.

1990: *I limiti dell'interpretazione*, Bompiani, Milano.

Elia, A.

1984: *Le verbe italien*, Schena-Nizet, Fasano-Paris.

Elia, A., Martinelli, M., D'Agostino, E.

1981: *Lessico e strutture sintattiche*, Liguori, Napoli.

Firenze, A.

1987: *Présentation des noms composés de l'italien, recherche d'un traitement formel*, Mémoires du D.E.A. d'Informatique Fondamentale, Paris, Université Paris 7.

Giry-Schneider, J.

1987: *Les prédicats nominaux en français*, Droz, Genève.

Godart, L.

1992: *L'invenzione della scrittura*, Einaudi, Torino.

Grishman, R.

1986: *Computational Linguistics. An Introduction*, Cambridge University Press, Cambridge (trad. it. *Linguistica computazionale*, Tecniche Nuove, Milano, 1988).

Gross, G., Vivès, R.

(éds.) 1986: "Syntaxe des noms", *Langue Française*, n. 69, Larousse, Paris.

Gross, M.

1968: *Grammaire transformationnelle du français. 1- Syntaxe du verbe*, Cantilène, Paris.

1975: *Méthodes en syntaxe, régime des constructions complétives*, Hermann, Paris.

1977: *Grammaire transformationnelle du français. 2- Syntaxe du nom*, Cantilène, Paris.

1989: "La construction de dictionnaires électroniques". *Annales des Télécommunications*, tome 44, n° 1-2, p. 4-19, Issy-les-Moulineaux/ Lannion: CNET. (<http://infolingu.univ-mlv.fr/>, menu Bibliographie, texte intégral)

1991: *Grammaire transformationnelle du français. 3 - Syntaxe de l'adverbe*, Maurice Gross et Asstril, Paris.

Guillet, A., Leclère, Ch.

(éds.) 1981: “Formes syntaxiques et prédicats sémantiques”, *Langage*, n. 63, Larousse, Paris.

1992: *La structure des phrases simples en français. 2. Constructions transitives locatives*, Genève-Paris, Droz.

Harris, Z. S.

1963: *Discourse Analysis Reprints*, La Haye, Mouton.

1968: *Mathematical Structures of Language*, Wiley, New York, (trad. fr. *Structures mathématiques du langage*, Dunod, Paris, 1971).

1970: *Papers in Structural and Transformational Linguistics*, Dordrecht, Reidel.

1976: *Noters du cours de syntaxe*, Le Seuil, Paris.

1988: *Language and Information*, Columbia University Press, New York.

Hornby, A. S., (éd.)

1989: *Oxford Advanced Learner's Dictionary*, Oxford University Press, Oxford.

Lanuzza, S.

1994: *Storia della lingua italiana*, Newton Compton, Roma.

Laporte, É.

1992: “Adjectifs en -ant dérivés de verbes”, in *Langue Française*, n. 96, Larousse, Paris.

1993: “Separating Entries in Electronic Dictionaries of French”, in *Sprache - Kommunikation - Informatik. Akten des 26 Linguistischen Kolloquiums. Poznan 1991*, Niemeyer, Tübingen.

Lepschy, C. G.

1989: *Sulla linguistica moderna*, Il Mulino, Bologna.

Lyons, J.

1968: *Introduction to Theoretical Linguistics*, Cambridge University Press, Cambridge (trad. it. *Introduzione alla linguistica teorica*, Laterza, Bari, 1971).

Martinet, A.

1960: *Éléments de linguistique générale*, Colin, Paris (trad. it. *Elementi di linguistica generale*, Laterza, Bari, 1966).

Matthews, P. H.

1973: *Morphology. An Introduction to the Theory of Word-Structure*, Cambridge University Press, Cambridge-London (trad. it. *Morfologia. Introduzione alla teoria della struttura della parola*, Il Mulino, Bologna, 1979).

Monteleone, M.

1989a: "Les expressions figées de l'italien: l'utilisation du verbe *fare*", in AA. VV., *Mémoires du D.E.A. d'informatique fondamentale 1989-1990*, Ceril-Université Paris 7, Paris.

1989b: "Comparazione delle espressioni idiomatiche dell'italiano, del *français normé* e del *français du Québec*. Un campione di trattamento lessicale e sintattico degli scarti", in Associazione Italiana di Studi Canadesi (a cura di), *Canada novecento/2. Rivista di Studi Canadesi*, n. 2-Anno 1989, Schena, Fasano.

1996: *Survey and Testing of Finite-State Methods for the Recognition of Agglutinated Sequences in Italian with the Aid of Lexical Information about Possible Combinations of Verbs and Clitics*, rapporto tecnico del progetto CEE Copernicus "Gramlex 621".

Monteleone, M., Musto, R.

1995: "Un esempio di applicazione dell'analisi harrisiana alla didattica delle lingue" in D'Agostino, E. (a cura di), *Tra sintassi e semantica. Descrizioni e metodi dell'elaborazione automatica della lingua d'uso*, ESI, Napoli.

Morvan, P.

1988: *Dictionnaire de l'informatique*, Larousse, Paris (trad. it. *Dizionario di informatica*, Gremese-Larousse, Roma, 1989).

Mounin, G.

1963: *Les problèmes théoriques de la traduction*, Gallimard, Paris (trad. it. *Teoria e storia della traduzione*, Einaudi, Torino, 1965).

1972: *Clefs pour la sémantique*, Seghers, Paris (trad. it. *Guida alla semantica*, Feltrinelli, Milano, 1983).

Orwell, G.

1949: *Nineteen Eighty-Four*, The Estate of Eric Blair (trad. it. *1984*, Mondadori, Milano, 1950).

Pantaleo, N., a cura di

1985: *Ideologia linguaggio potere: 1984 di G. Orwell*, Adriatica Editrice, Bari.

Peirce, Ch. S.

1931-1935: *Collected Papers*, The Belknap Press of Harvard University Press, Cambridge - Mass. (trad. partielle dans *Semiotica. I fondamenti della semiotica cognitiva*, Einaudi, Torino, 1980).

Pinker, S.

1994: *The Language Instinct. How the Mind Creates Language*, HarperPerennial, New York (trad. it. *L'istinto del linguaggio. Come la mente crea il linguaggio*, Arnoldo Mondadori Editore, Milano).

Perniola, M.

1994: *Il sex-appeal dell'inorganico*, Einaudi, Torino.

Popper, K.R.

1934: *Logik der Forschung*, Wien (trad. it. *La logica della scoperta scientifica*, Einaudi, Torino, 1970).

1984: *Auf der Suche einer bessern Welt. Vorträge und Aufsätze dreissig Jahren*, R. Piper GmbH & Co. KG, München (trad. it. *Alla ricerca di un mondo migliore*, Armando, Roma, 1989).

1991: *Scienza e filosofia*, Einaudi, Torino.

1992: *La lezione di questo secolo*, Marsilio, Venezia.

1994: *Alles Leben ist Problemlösen. Über Erkenntnis, Geschichte und Politik*, Karl Popper Estate, South Croydon, Surrey (trad. it. *Tutta la vita è risolvere problemi*, Rusconi, Milano, 1996).

Popper, K.R., Condry, J.

1996: *Cattiva maestra televisione*, Giancarlo Bosetti, Roma.

Renzi, L., a cura di

1988: *Grande grammatica italiana di consultazione. I. La frase. I sintagmi nominale e preposizionale*, Il Mulino, Bologna.

Sapir, E.

1921: *Language. An Introduction to the Study of Speech*, Harcourt, Brace & Company, New York (trad. it. *Il linguaggio. Introduzione alla linguistica*, Einaudi, Torino, 1969).

Saussure, F. de

1916: *Cours de linguistique générale*, Payot, Lausanne-Paris (trad. it. *Corso di linguistica generale*, Laterza, Bari, 1967).

Schank, R.

1984: *The Cognitive Computer. On Language, Learning and Artificial Intelligence*, Wesley Publishing Company, Inc., Reading, Addison (trad. it. *Il computer cognitivo. Linguaggio, apprendimento e Intelligenza Artificiale*, Giunti-Barbèra, Firenze, 1989).

Sensini, M.

1990: *La grammatica della lingua italiana*, Mondadori, Milano.

Serianni, L. (con la collaborazione di) Castelvechi, A.

1988 *Grammatica Italiana. Italiano comune e lingua letteraria. Suoni, forme, costrutti*, UTET, Torino.

Silberztein, M.

1993: *Dictionnaires électroniques et analyse automatique de textes. Le système INTEX*, Masson, Paris.

1997: *Manuel de référence du logiciel INTEX 3.4*, Université Paris 7-LADL, Paris.

Tekavcic, P.

1980: *Grammatica storica dell'italiano. Vol. II: Morfosintassi*, Il Mulino, Bologna.

Vietri, S.

1985: *Lessico e strutture sintattiche*, Liguori, Napoli.

2001: *Navigare nei testi. Applicazioni in linguistica computazionale*, Editoriale Scientifica, Napoli.

Wittgenstein, L.

1922: *Tractatus logico-philosophicus*, Routledge & Kegan Paul, London (trad. it. Einaudi, Torino, 1964).

1953: *Philosophische Untersuchungen*, Basil Blackwell, Oxford (trad. it. *Ricerche filosofiche*, Einaudi, Torino, 1967).